THIS WEEK

EDITORIALS

Animal farm

Europe's policy-makers must not buy animal-rights activists' arguments that addiction is a social, rather than a medical, problem.

Drug addiction is a disease. Images of the brains of addicts show alterations in regions crucial to learning and memory, judgement and decision-making, and behavioural control. Drugs imitate natural neurotransmitters, resulting in false or abnormal messages being sent around neural circuits. The brain's central reward system is overstimulated and flooded with dopamine. The brain adapts to this flood by turning down its ability to respond to dopamine — so addicts take more and more of the drug to push dopamine levels higher.

Changes in other reward-system neurotransmitters such as glutamate can impair cognitive function. And the triggering of subconscious memory systems leads to conditioning, so environmental cues such as particular people or places set off uncontrollable cravings.

None of that is particularly controversial, at least among scientists. So why do a growing number of politicians in Europe want to curtail research into addiction? Why would they deny their constituents the hope that they or their loved ones might one day be helped with the terrible burden of this disease?

The answer is a troubling new front in the long battle over the use of animals in research (see page 24). Campaigners opposed to animal research have targeted addiction as the soft underbelly of political support for such work. Addiction is a social problem, they argue, not a medical one. And social problems are not solved by science, or by research on animals.

That is a seductive message for politicians. Care and compassion for drug addicts is rarely a vote-winner. Care and compassion for animals is a sure thing. Many voters believe that funds are best focused on crushing drug barons and locking up dealers. Many also believe that addicts are at best weak-minded, at worst evil, and have only themselves to blame if their drug habits kill them. If the science of addiction can be questioned, then why bother pursuing medical cures based on scientific research?

(For a taste of the muddled thinking on offer here, search the Internet for a recent 'debate' on addiction featuring the journalist Peter Hitchens and the actor Matthew Perry, broadcast by the BBC's current-affairs programme *Newsnight*.)

DANGEROUS DECREE

Flawed, unscientific thinking on addiction has already produced a decree in Italy, expected to become law next month, which bans the use of all animals in addiction research — despite vociferous objections from the scientific community. The dozen or so Italian groups working in this area will have three years to phase out their research, and other scientists hoping to develop a drug for any brain-related disorder, from anxiety to migraines, will no longer be able to generate safety data required by regulatory agencies on their addictive potential in animals. In Belgium, the government is rushing through legislation that would ban addiction research using monkeys — again in the face of objections from scientists.

Let's be clear: research using animals has been central to our understanding that drug addiction is a chronic, relapsing disease that changes the structure and function of the brain; that an individual's genetic make-up accounts for around half of the vulnerability to addiction; and that environmental factors are crucial in precipitating addictive behaviours in the vulnerable. Environmental factors include stress at critical developmental stages, from the womb, through early childhood, to adolescence. Without animal research — including work on primates, whose brains are most like our own — it will not be possible to go further and discover exactly how the neural circuitry in an individual brain is shaped by interacting genetic and environmental forces. It is an extremely tough nut to crack — but it must be cracked.

> ***"Winning the war against the misuse of drugs requires us to address demand as well as supply."***

Research using animals is, rightly, a perennially sensitive issue. But to claim that animals may not be used specifically for addiction research is to define those affected by the disorder as less worthy of care and concern than those with other disorders. Politicians cannot choose to ignore scientific evidence and then claim that they do not know that addiction is a disease. Animal-rights campaigners have unleashed a dangerous argument. It must be stopped in its tracks — and quickly.

Winning the war against the distress and damage caused by misuse of drugs requires diverse approaches that address demand as well as supply. It may not seem intuitive to those witnessing the misery and violence around the drug world — in the United States alone, illicit drug use costs more than US$190 billion a year in crime, increased health-care costs and lost productivity — but it is likely that demand can be reduced by developing treatments for the self-destructive cravings that drive drug addiction. Given the technical tools now available for looking deep inside the brain, there is realistic hope that such treatments will emerge from research in the coming decades.

The work must continue. Europe should look to the United States and to inspirational figures such Nora Volkow, head of the US National Institute on Drug Abuse in Bethesda, Maryland, who regularly testifies on the science of addiction to the US Congress to justify the institute's research budget.

Volkow — a neuroscientist born in Mexico, a country blighted by drug wars — has the scientific clarity of vision, and the relentless patience, to be able to argue for the promise of research effectively year in, year out. Such wisdom also exists in Europe, but politicians too frequently ignore it.

Volkow is the great-granddaughter of Leon Trotsky, the Russian revolutionary who was famously assassinated in 1940 — in the family home in Mexico in which Volkow herself grew up. She fights for a different cause: rational drug politics. Her European counterparts are fighting too. All governments must pay attention.

Diseases can be cured. People affected can be helped by science and research — and yes, by the use of animals. Addiction is no different. ■

Invisible borders

UK immigration rules are perceived as being tougher than they really are.

For an international business such as science, there is no true immigration — just movement. National borders are notional; passports indicate where someone was born, not where they belong or where they can do the most good.

Politicians, and much of the public, do not see immigration like that. They see strain on services and overcrowded job markets and, in a few cases, some use these legitimate concerns as thin masks for prejudice, xenophobia and racism. As economic difficulties bite, governments frequently promise to 'get tough' on immigration, often turning a blind eye to the benefits of an influx of people as they do so.

In 2010 the warnings were stark. The United Kingdom's new, firmer stance on immigration could "spell disaster for UK science", warned the Royal Society of Chemistry. Nobel laureates voiced similar concerns in the national press. The Campaign for Science and Engineering (CaSE) mobilized to keep the United Kingdom 'open for business', with calls to give priority to visa applications from overseas scientists.

Nature too spoke out about the risk posed by crude measures to curb immigration, which could potentially scupper the ability of some of Britain's leading research laboratories to recruit the best people (see *Nature* **468,** 346; 2010).

To the UK government's credit, it heeded the warnings and made exceptions for scientists, loosening the rules to grant them entry under circumstances that would cause other migrants from different professions to be turned away. A new scheme for 'exceptionally talented' scientists and artists was created. Some 700 scientists and 300 artists a year would be allowed in. Problem solved?

Apparently not. The United Kingdom is again in the throes of a political debate about the benefits and problems of immigration, and science lobby groups are again worried that researchers will be caught in the crossfire. As we report on page 14, these groups recently badgered the Home Office at a meeting on the subject, and some made the same doom-laden predictions.

Unfortunately, the campaigners are on less solid ground this time. The exceptional-talent route has mainly been exceptional in its underuse: the latest figures show that by June 2013 only 89 people (both scientists and artists) had used it. Indeed, one researcher told *Nature* that his visa is so rare that "when I re-enter the UK, the border staff always comment 'Oh, I've never seen one of those'".

Hundreds of places on the scheme remain unfilled. And other concessions remain. In fact, scientists are in a better position than just about anyone else who wishes to move to and work in the United Kingdom, with the possible exception of international soccer stars. And even there, similar rhetoric about the effect on domestic talent has led to footballers being recently refused work permits.

"Scientists are in a better position than just about anyone else who wishes to move to and work in the United Kingdom."

The situation is not perfect — far from it. Many academics have justifiable gripes with the UK Border Agency and its visa processes — a postdoc forced out of the country near-penniless perhaps, or an eminent colleague scheduled to deliver a keynote speech at a conference turned away. Such difficulties are common throughout most of the world. The United States is currently grappling with how to keep happy the thousands of scientists who are unable to obtain green cards for permanent residency every year.

'Highly skilled' migrants are usually singled out for praise when politicians confront immigration, but the subject is a notorious minefield and is hard for politicians to navigate with rational arguments. Image, perceived approach and rhetoric about being 'tough' buy popularity here perhaps more than in any other political sphere — whatever the evidence may say.

After the recent meeting with the Home Office, CaSE said that such "messaging" from the UK government about making it harder for immigrants could itself deter leading scientists from coming. That could explain, for example, why piles of the exceptional-talent visas remain unused in the drawers of the Border Agency. But seen another way, the government is making it possible for scientists to come, whereas it is the campaigners who organize open letters and give media-friendly briefings about how hard it is for them to do so. Pressure groups have one weapon — pressure — and it is one that can be as crude as any political rhetoric. If messaging is the problem, then the campaigners must be cautious about the message that they themselves send.

There are real problems with the movement of scientists across borders, and campaigners are right to highlight them. *Nature* will continue to press for such obstacles to be removed — the real and the rhetorical. ■

Trick of the light

The Amazon doesn't absorb extra carbon in the dry season after all. It can become a carbon source.

Budding biologists learn early the apparently simple holy trinity of ingredients for photosynthesis: carbon dioxide, water and light. In truth, the equation is a little more complicated than that, and when photosynthesis proceeds on a truly massive scale, these complications can have huge implications.

Take, for example, the world's largest mass of concentrated photosynthesis: the Amazon rainforest of South America. Scientists have long struggled to work out whether the rate of photosynthesis there is controlled by the available amount of water or of sunlight. (Over seasonal timescales, that is — on a 24-hour cycle, it is controlled by the availability of sunlight.)

The uncertainty was triggered by a surprising result from satellite images, which seemed to show that Amazon forests became greener during the dry season, and greenest of all during years of severe drought such as 2005 (S. R. Saleska *et al. Science* **318,** 612; 2007). More green means more photosynthesis, so this result suggested that it was the availability of light, and not water, that was the controlling factor. Clear skies and sunny weather were more important than moisture in the soil.

In a study published on *Nature*'s website today (D. C. Morton *et al. Nature* http://dx.doi.org/10.1038/nature13006; 2014), researchers show that this is, literally, an illusion. The forest does not become greener during dry periods at all. It just looks that way when the sensor and the Sun are both in the south of the sky. It is not photosynthesis that drives the apparent greening of the forest at such times, but a lack of shadow.

The finding drags attention away from the importance of light in the Amazon's photosynthesis equation, and towards the need for water. But what of the third point of the triangle, carbon dioxide? There is uncertainty there too: this time over whether in years of drought, the trees will switch from being a net carbon sink to a source, which could worsen global warming. A second study of the Amazon, on page 76, offers the latest data on this debate, and the news is not good. Fire and drought can indeed make the Amazon a net source of atmospheric carbon — whatever colour it is at the time. ■

➲ NATURE.COM
To comment online, click on Editorials at: go.nature.com/xhunqv

WORLD VIEW *A personal take on events*


YULAN LI

China must protect high-quality arable land

Figures from a national survey of land use seem positive, but the effort exposed some worrying trends, says **Xiangbin Kong**.

According to the Chinese government, China needs a minimum of 120 million hectares of arable land to feed its people. That is the 'red line' for food security that officials have pledged to protect.

So it may seem like good news that the most recent comprehensive survey of national land use in China has reported a healthy surplus — some 135 million hectares of the country are classed as planted with crops — rice paddy fields, irrigable land and dry farms. Simultaneously, total grain production hit a record 602 million tonnes in 2013, after a decade of continuous growth.

I fear these figures are not as positive as they seem. Moreover, I worry that they may create a false sense of security and encourage policy-makers to relax efforts to protect China's arable land. We should not be misled by the superficial surplus. The story is not so simple. Although the quantity of arable land in China seems healthy, there are serious concerns about its quality — and of its ability to supply future generations with enough food.

I was involved in the land-use assessment — the second National Land Resource Survey of China — and am pleased to see the results published and discussed. The survey completed its work in 2009, but the previous central government declined to publish the results, because members did not agree with the findings of the first such survey, finished in 1996.

This is not unusual. Surveys to classify large areas of remote land are difficult. A study published in December suggested that the area of cropland abandoned since 1990 in western Russia, Belarus and Ukraine has been severely underestimated.

When President Xi Jinping came to power in 2012, he investigated the discrepancy in the Chinese figures and decided that the results of the second survey were robust, because they are based on high-resolution remote sensing and backed up by investigation on the ground. He authorized China's land and resources ministry (MLRC) to release the results at the end of 2013. At a press conference on 30 December, officials from the MLRC pledged to continue to protect arable land, and reinforced their commitment to the food security red line.

Beyond the headline figures, there are some worrying trends. Although the overall area of arable land has increased in the time between the two surveys, the quality of the land, and so its suitability, has decreased. Some 3 million hectares of high-quality arable land and some 1 million hectares of paddy land have been built on or converted to urban use in just over a decade. More than 3 million hectares have been contaminated with pollution. The effects were shown starkly last year, when heavy metals such as cadmium appeared on the tables of restaurants as a result of rice being planted in polluted fields in Hunan province.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/nzytqp

Arable land lost to development and contamination is frequently replaced by marginal and lower-quality alternatives — although land surveys such as ours do not distinguish between them. Of the land identified as arable in the latest figures, more than 4 million hectares in the southwest of the country are high in the mountains. And almost 6 million hectares are in converted forest and grassland in the north, an ecologically fragile flood zone. Broadly, there has been a shift from growing crops in China's warm and humid south to the less suitable cold and water-limited north.

The amount of available land has peaked. There is no spare high-quality arable land that can be cultivated as existing farmland is lost to development. Further conversion of grassland and forest produces low-grade alternatives, at great ecological cost. Rather than signalling security, the new land-use figures show that China is overusing its remaining high-quality arable land. It is growing more food on less land, a situation that leaves little scope for expansion — and little in reserve as water shortages reduce yields in the north still further.

> CHINA IS GROWING MORE FOOD ON **LESS LAND,** A SITUATION THAT LEAVES **LITTLE SCOPE** FOR EXPANSION.

China needs to act to preserve its remaining high-quality arable land by classifying tracts of land as for permanent arable use, particularly in the southeast and in the suburbs of big cities. Restrictions should be put on development there, and greater efforts made to prohibit the agricultural conversion of marginal land in the north. Together, this would slow the agricultural shift towards the north and buy China some time.

China should also rethink its existing protection policy for arable land, which contributes to the problem because its programmes focus only on individual administrative regions. Instead, China should set aside crop production 'priority zones' at a national, provincial and county level on the basis of the arable land's potential grain productivity and distribution.

These zones should introduce protection for other types of farmland, increase the subsidies paid to households that increase crop production per hectare, and make available funds for reclamation and restoration of degraded land to produce crops. This is a key point. China must work harder to improve the quality of low- and medium-grade arable land, which the country will increasingly rely on to feed itself. Better rural roads, forest management, and more irrigation canals and ditches are less newsworthy than headline announcements about record crop production, but, in the medium and long term, they will be more useful. ■

Xiangbin Kong *is a land-use scientist at the College of Resources and Environment, China Agricultural University, Beijing.*
e-mail: kxb@cau.edu.cn

RESEARCH HIGHLIGHTS

Selections from the scientific literature

BIOCHEMISTRY

Prion strings pictured on cells

For the first time, researchers have captured images of prions — proteins that can misfold and spread, causing neurodegeneration — in living cells. The images show the proteins residing on the cell surface in strings and webs.

Albert Taraboulos at the Hebrew University in Jerusalem and his colleagues used antibodies that react with a subset of the misfolded proteins to visualize the prions in cultured mouse cells and brain tissue under a fluorescence microscope. The team found prion strings up to five micrometres long that remained stable on the cell surface for several hours.

This anchoring provides insight into how misfolded prions interact with cells and can resist degradation, the authors say.

J. Cell Biol. **http://dx.doi.org/10.1083/jcb.201308028 (2014)**

Y. NUI *ET AL.*/CELL PRESS

BIOTECHNOLOGY

CRISPR makes modified monkeys

Researchers have used precise gene-editing techniques to generate genetically modified monkeys.

Previous models of human disorders in monkeys were created using viruses to transfer genes, but this method lacks the precision needed to modify specific gene sequences. Xingxu Huang at Nanjing University in China and his colleagues turned to the CRISPR–Cas9 system, which uses a customizable RNA fragment to guide the DNA-cutting enzyme Cas9 to a specific site. The team altered the genome in one-cell-stage embryos of cynomolgus monkeys (*Macaca fascicularis*). This resulted in the birth of twins (pictured) with mutations in two target genes: *Ppar-γ*, which is involved in regulating metabolism; and *Rag1*, which is involved in immune function.

The results pave the way for producing primate models with specific mutations that more closely mimic human diseases.

Cell **http://doi.org/q93 (2014)**

For a longer story on this research, see go.nature.com/327cbd

ASTROPHYSICS

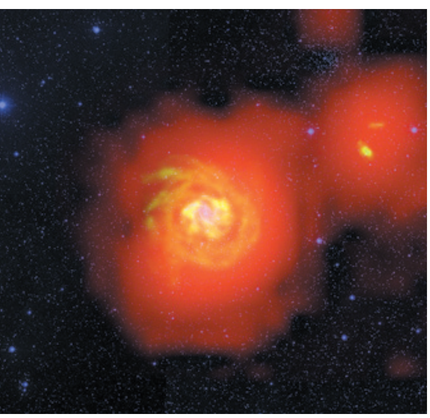
Hydrogen river could fuel stars

The discovery of a faint filament of hydrogen gas streaming across space could help to explain how some galaxies maintain their pace of star formation.

D. J. Pisano from West Virginia University in Morgantown used the Robert C. Byrd Green Bank Telescope to identify a river of hydrogen connecting the galaxy NGC 6946 (**pictured**) with its neighbours. Pisano suggests that the filament could be the first observation of a 'cold flow', a stream of diffuse gas from intergalactic space that has long been theorized to be a source of fuel for star formation, and that is invisible to most telescopes.

Alternatively, the hydrogen could have been drawn out during a close encounter between NGC 6946 and its neighbours. Future galaxy surveys should confirm the source of this hydrogen stream.

Astronomical J. 147, **48 (2014)**



D. J. PISANO (WVU)/B. SAXTON (NRAO/AUI/NSF)/PALOMAR OBSERVATORY, SPACE TELESCOPE SCI. INST. 2ND DIGITAL SKY SURVEY (CALTECH)/WESTERBORK SYNTHESIS RADIO TELESCOPE

ARCHAEOLOGY

Britain's Anglo-Saxons were local

Anglo-Saxons succeeded the Romans in Britain during the early fifth century, probably through cultural adoption by local individuals rather than through invasion by Germanic people.

Susan Hughes at the US Navy in Silverdale, Washington, and her team analysed the tooth enamel of 19 individuals from an early Anglo-Saxon cemetery in southern England, and measured the levels of oxygen and strontium isotopes in the teeth. These levels are determined by the water and food consumed by the individual. The researchers found that the isotope ratios matched those of the surrounding water and soil, suggesting that most of the people were local to that area. One individual seemed to be an immigrant from the

European continent.

The team says that its findings support the idea that Britain's first Anglo-Saxons were locals who rapidly shifted cultures after the fall of Roman Britain.

J. Arch. Sci. 42, **81–92 (2014)**

ANIMAL BEHAVIOUR

Lizards socialize to thrive

Social isolation in early life could impair the development of reptiles, according to a study of chameleons.

Social behaviour is well documented in mammals and birds, but it is not so firmly corroborated in cold-blooded vertebrates. Cissy Ballen and her colleagues at the University of Sydney in Australia compared the social interactions of veiled chameleon (*Chamaeleo calyptratus*) hatchlings raised in isolation with those raised in a group setting. The authors found that socialized lizards were less submissive, displayed brighter and more saturated colours when encountering new chameleons, and captured food more quickly than did lizards raised in isolation.

The findings add to evidence challenging the conventional view that reptiles are capable of only simple social behaviour.

Anim. Behav. **http://doi.org/q9h (2014)**

EVOLUTIONARY BIOLOGY

Night life fosters foul sprays

Carnivores that spray foul anal secretions might have evolved this ability in response to night-time predation from other mammals.

Theodore Stankowich at California State University in Long Beach and his colleagues looked at the behaviour of 181 species of carnivorous mammals and their predators. The authors found that carnivores are targeted mainly by other mammals at night and by birds of prey during the day. Animals that are active during the day are more likely to develop tight-knit social groups that are better at detecting and warding off predators.

Nocturnal animals cannot rely on early visual detection and instead use short-range defence systems such as noxious sprays, which are more effective against other mammals than against birds.

Evolution **http://doi.org/q9w (2014)**

NEUROSCIENCE

Pruning problems alter brain wiring

Abnormal pruning of neuronal connections might stall brain maturation, resulting in reduced brain connectivity and even behaviours linked to disorders such as autism.

Cornelius Gross at the European Molecular Biology Laboratory in Monterotondo, Italy, and his colleagues studied mice that were engineered to have fewer microglia — non-neuronal brain cells that trim back synapses, or neuronal connections, during brain development. These animals had fewer synapses between neurons and decreased connectivity between brain regions, and seemed to be less social in behavioural tests.

Microglia and synaptic pruning are important for normal brain development, and problems with this pruning could lead to neurodevelopmental disorders, the authors say.

Nature Neurosci. **http://doi.org/rbf (2014)**

ASTRONOMY

How big galaxies died fast

Astronomers have worked out the origin of giant galaxies that seemed to have fizzled early in the Universe's history, just three billion years after the Big Bang.

To find out how massive elliptical galaxies became so big and stopped forming stars so quickly, Sune Toft of the Niels Bohr Institute in Copenhagen and his colleagues compared samples of these dead galaxies and an earlier generation of star-forming ones observed with the Hubble, Herschel and Spitzer space telescopes. The authors conclude that earlier, gas-rich galaxies merged, kicking off intense star formation that rapidly used up all the gas, resulting in the large, burnt-out galaxies.

Astrophys. J. 782, **68 (2014)**

COMMUNITY CHOICE

The most viewed papers in science

CLIMATE CHANGE

United States tops warming list

HIGHLY READ on iopscience.iop.org 29 Dec–28 Jan

The United States is the largest national contributor to global climate warming, followed by China, Russia, Brazil and India.

Damon Matthews and his colleagues at Concordia University in Montreal, Canada, analysed the national emissions of greenhouse gases and aerosols, including those from land use, between 1800 and 2005. They calculated that a total warming of 0.7 °C occurred during this period, and that more than 21% of this total is linked to the United States. China and Brazil exceed the United States slightly in terms of their contributions from land-use activities, such as deforestation and agriculture, but the high level of cumulative US fossil-fuel use makes the country the biggest contributor overall.

Among the major emitters, the United Kingdom and the United States top the rankings on a per capita basis, with contributions that are more than ten times higher than those of either China or India.

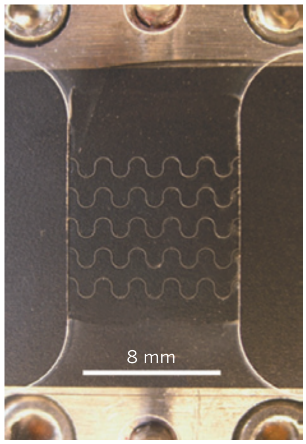
Environ. Res. Lett. 9, **014010 (2014)**

MATERIALS

Tiny cracks toughen up glass

Glass etched with intricate micropatterns is much tougher than normal glass, report Francois Barthelat and his colleagues at McGill University in Montreal, Canada.

The researchers were inspired by natural materials such as tooth enamel and nacre in mollusc shells, which are stiff and hard, but not brittle. In these structures, cracks are unable to spread rapidly because they are forced to travel along tortuous or interlocking channels that are criss-crossed by proteins holding the structure together. The researchers etched similar patterns into glass (**pictured**) and filled in the gaps with shock-absorbent polyurethane, creating a material that is 200 times tougher than standard glass.



M. MIRKHALAF *ET AL.*/*NATURE COMMUN.*

The approach could be used to make brittle materials such as ceramics shatter-resistant.

Nature Commun. 5, **3166 (2014)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

Dengue control

Panama has joined a handful of nations trying to combat dengue fever using genetically modified mosquitoes developed by Oxitec, a biotechnology firm in Oxford, UK (see go.nature.com/tht55x). The company announced on 28 January that the Panamanian government had approved open-field trials of the insects, engineered to be sterile, as a means of suppressing wild populations of the dengue-carrying mosquito (*Aedes aegypti*). At the end of 2013, Panama's health minister declared that the country was experiencing a dengue epidemic.

Oil pipeline

On 31 January, the US Department of State released its final environmental assessment of the proposed Keystone XL pipeline — a controversial project that would link oil sands in Alberta, Canada, to refineries along the Gulf of Mexico. Environmentalists have argued that the pipeline would increase carbon emissions by facilitating fuel production. But the agency concluded that approval or denial of the pipeline is "unlikely to significantly impact" production or consumption of the oil, which could travel by alternative routes. The White House must now decide whether to proceed with the project.

Food foundation

The sweeping US$956-billion 'Farm Bill', passed by the US House of Representatives on 29 January, includes some research support among its wider measures regulating food-stamp payments and farmer subsidies. The bill, formally known as the Agricultural Act of 2014, authorizes the creation of a non-profit corporation called the Foundation for Food and Agriculture Research. It calls for an initial $200 million for the foundation to support research in areas such as plant and animal health, nutrition and renewable energy. The Senate is expected to pass the bill this week.

RICHARD ELLIS/GETTY

Butterfly migration hits historic low

Fewer monarch butterflies migrated across North America in 2013 than in any previously recorded year, according to a report released on 29 January by conservation group the WWF. Surveys of forested regions in Mexico, where monarch butterflies (*Danaus plexippus*) hibernate from November to March, found the creatures occupying only 0.67 hectares of land — a 44% drop from the previous year, and the smallest area since surveys began in 1993. Area occupied is used as an indicator of population size. Changes in land use and extreme climate conditions along the roughly 4,000-kilometre migration route from Canada to Mexico have contributed to the decline, as has deforestation of hibernation sites, says the WWF. Use of agricultural herbicides has reduced the availability of milkweed, a key food source.

Reef waste dump

The authority in charge of Australia's Great Barrier Reef has approved a controversial dumping of dredging waste in the marine park around the immense coral edifice. The Great Barrier Reef Marine Park Authority says that "strict environmental conditions" will be in place for the dumping of up to 3 million cubic metres of spoil originating from the expansion of the Abbot Point coal port. But environmental groups say that the move will endanger the ecosystem and threatens the reef's status as a UNESCO World Heritage Site.

FUNDING

Stem-cell genomics

Researchers from seven Californian institutions have bagged a US$40-million grant to establish a centre that will apply large-scale genetics studies to stem-cell research. The award from the California Institute for Regenerative Medicine in San Francisco, announced on 29 January, will support a Center of Excellence in Stem Cell Genomics led jointly by Stanford University in Palo Alto and the Salk Institute for Biological Studies in San Diego. The selection process raised protests from other applicants, who questioned departures from review procedures used in previous grant cycles.

RESEARCH

Cancer genetics

The US National Cancer Institute in Bethesda, Maryland, has launched one of the first trials to assess whether

cancer treatments that are tailored to individual genetic profiles are more beneficial for patients than non-targeted treatments. The Molecular Profiling based Assignment of Cancer Therapeutics (M-PACT) study, announced on 30 January, will screen tumours from 180 patients for mutations in 20 genes that are known to affect treatment. Half of the patients will then receive therapy that is customized for their specific mutations, and half will receive non-customized therapy. The findings are expected to be reported in 2017.

Lab-animal reforms

Responding to criticisms, Imperial College London on 31 January unveiled a plan for "wholesale reform" of the ethical review and governance of its animal research. Last year, the university underwent an independent review after an undercover investigation by anti-vivisectionists produced allegations of malpractice (see *Nature* http://doi.org/rbd; 2013). See go.nature.com/7vqf2t for more.

PEOPLE

Science leader

Chemist Geraldine Richmond (**pictured**) has been chosen as the next president-elect of the American Association for the Advancement of Science

KELLY JAMES

(AAAS) in Washington DC. Richmond, who is a professor at the University of Oregon in Eugene, studies the chemistry of surfaces and interfaces. She is a member of the US National Academy of Sciences and the founder and chair of the Committee on the Advancement of Women Chemists, an organization that supports female scientists and engineers. She will take over as AAAS president in February 2015.

BUSINESS

Neglected diseases

Global health advocates expressed dismay last week over news that pharmaceutical giant AstraZeneca is ending research on treatments for tuberculosis, malaria and neglected tropical diseases. In 2012 the company, which has its headquarters in London, joined a coalition to eradicate neglected tropical diseases, which affect 1.4 billion people worldwide. Geneva-based advocacy group Médecins Sans Frontières called the latest move "discouraging" and highlighted the need to combat neglected diseases that afflict the world's poorest people (see *Nature* **505,** 142; 2014).

Data partners

Pharmaceutical giant Johnson & Johnson announced on 30 January a new partnership with Yale University in New Haven, Connecticut, to share data from the company's clinical trials. The Yale University Open Data Access project will serve as an independent body to review and manage requests from researchers seeking anonymized clinical-trial data from the company, of New Brunswick, New Jersey. The move follows initiatives to promote clinical data sharing and transparency in the United States and in Europe (see *Nature* **505,** 131; 2014).

School suspension

Educational company Coursera has blocked access to services for students from Cuba, Iran and Sudan. The firm, which is based in Mountain View, California, specializes in massive open online courses (see *Nature* **495,** 160–163; 2013). Citing US export regulations that restrict services to sanctioned nations, Coursera said on 28 January that it had begun blocking users from logging on to its website from IP addresses in affected countries. Access for students from Syria was initially revoked, but was reinstated after the company learned of a regulatory exception.

Drilling deferred

Oil company Royal Dutch Shell has shelved its 2014 drilling programme off the coast of Alaska. Speaking to investors on 30 January, chief executive Ben van Beurden cited a 22 January court ruling that the US government did not properly assess the potential environmental impacts of offshore drilling in the Chukchi Sea (see *Nature* **505,** 590; 2014). "The lack of a clear path forward means that I am not prepared to commit further resources for drilling in Alaska in 2014," he said.

COMING UP

6–9 FEBRUARY
The Molecules and Materials for Artificial Photosynthesis conference in Cancún, Mexico, highlights the latest research in molecular catalysts, solar cells and nanomaterials for energy conversion and storage.
go.nature.com/asfpi5

12–15 FEBRUARY
Researchers gather in Marco Island, Florida, for the annual Advances in Genome Biology and Technology meeting. Topics include prospects for next-generation sequencing in cancer treatment, and the use of genomic tools to study neural circuits.
go.nature.com/dmkkmp

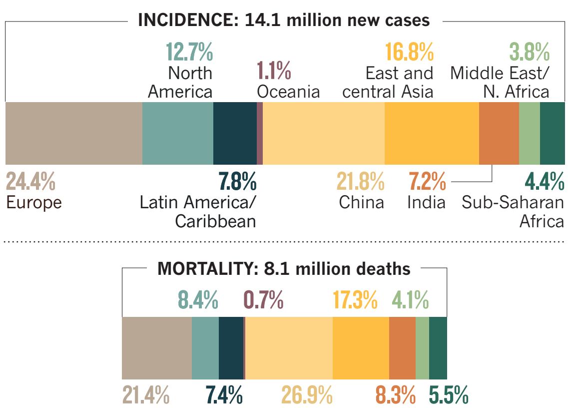
TREND WATCH

By 2025, there will be more than 20 million new cancer cases per year, compared with 14.1 million in 2012, according to the *World Cancer Report 2014*, released on 3 February by the World Health Organization's International Agency for Research on Cancer. Demographic changes and increased life expectancy are responsible, the report says. The greatest impact will be on low- and middle-income countries, noted Margaret Chan, director-general of the World Health Organization.

SOURCE: IARC WORLD CANCER REPORT 2014

THE BURDEN OF CANCER

Asia experienced around 46% of global new cancer cases in 2012, but 53% of cancer mortalities.



NATURE.COM
For daily news updates see:
www.nature.com/news

IMMIGRATION UK's welcome to foreign talent bears touch of frost **p.14**

PUBLISHING Elsevier opens its journals to text mining, subject to terms **p.17**

VITICULTURE World's premier grapevine gene bank faces eviction **p.18**

CLIMATE CHANGE Grim forecast for future Winter Olympics **p.20**

DAVID DUPREY/AP

There is little evidence that bite marks on a crime victim's skin allow reliable identification of the perpetrator.

CRIMINOLOGY

Faulty forensic science under fire

US panels aim to set standards for crime labs.

BY SARA REARDON

For 19 years, Gerard Richardson sat in prison in New Jersey wondering how forensics experts had got his case so wrong. His conviction for a 1994 murder was based on a bite mark on the victim's body that seemed to match his own teeth; it was the main physical evidence linking him to the crime. Last year, he was exonerated when DNA taken from the same bite mark turned out not to be his. According to the Innocence Project in New York, which tracks wrongful convictions, more than half of DNA exonerations involve faulty forensic evidence from crime labs and unreliable methods such as bite-mark analysis.

Cases such as Richardson's are one reason why the US Department of Justice and the National Institute of Standards and Technology (NIST) have now created the first US national commission on forensic science. The panel of 37 scientists, lawyers, forensics practitioners and law-enforcement officials met for the first time this week in Washington DC, and aims to advise on government policies such as training and certification standards. In March, NIST will begin to set up a parallel panel, a forensic-science standards board that will set specific standards for the methods used in crime labs.

For many scientists, this hard look at forensic science comes none too soon. "The broad objective is to put the science into forensic science so it can legitimately have the name," says commission member Stephen Fienberg, a statistician at Carnegie Mellon University in Pittsburgh, Pennsylvania. In 2009, the National Research Council (NRC) released a damning report criticizing US forensics practices. According to the report, nearly every analytical technique, from hair-sampling methods to those used in arson investigation, is unreliable, with too much variability in test results. Only DNA evidence escaped condemnation.

In addition, the NRC was concerned about forensics lab training. In 2009, only 60% of publicly funded crime labs employed a certified examiner. And the report called for standards to ensure that all labs evaluate evidence in the same way. Very often, it said, two labs analysing evidence from a crime scene will come up with different results using the same method.

The NRC offered a list of fixes, including the creation of a government agency with regulatory power and a research budget. Much like the NRC, the commission is only an advisory body that will offer expert opinions. But by having the ear of the US Attorney General, who can order changes in federal-agency practices, the national commission could be influential, says John Butler, a forensic geneticist at NIST and the commission's vice-chair. The commission will meet and produce recommendations until April 2015, although Butler says that its remit may be extended.

The two panels' recommendations will not directly affect practices in state and local labs, which handle more than 90% of forensics needs. But their visibility could cause recommended standards to trickle down. If that does not work, the federal government could withhold grants to labs that do not conform to new standards, or limit access to federal DNA databases.

Even in DNA collection, there are discrepancies between standard practices in federal, state and individual labs. The FBI, for ▶

instance, records 13 specific base-pair locations, or loci, from DNA samples in its national database, to ensure that false matches do not occur. But in 2008, the San Francisco Police Department in California used a 30-year-old, low-quality DNA sample from a murder case to convict a 70-year-old man who was listed in its state database — even though only five loci were matched. In a database the size of California's, matching based on these five loci would identify an innocent person one-third of the time.

Even good standards and best practices do not mean that a technique is solid, says Fienberg. Trained polygraph operators, for instance, can obtain consistent test results, but whether the machines accurately detect lies is highly uncertain. Many law-enforcement agencies still use the technique, even though a 2003 NRC report found it to be unreliable.

"The fundamental issues with forensic science can be solved by fixing the science," says Suzanne Bell, a forensic chemist at West Virginia University in Morgantown. Bell says that the field needs more research funding. In 2012, the National Institute of Justice funded just US$5 million in basic forensic-science research.

The value of certain techniques is often overstated in court cases, says Simon Cole, who studies the history of science in the criminal justice system at the University of California, Irvine. Fingerprint comparison, for instance, is often presented as an exact science, but researchers have only recently begun to study just how well people can do the matching. A 2011 study found that professional examiners matched two fingerprints incorrectly once in every 1,000 times, and missed a correct match 7.5% of the time (B. T. Ulery *et al. Proc. Natl Acad. Sci. USA* **108,** 7733–7738; 2011). Cole would like the standards board to define a 'match' precisely, and to assess the extent to which different methods yield different results.

The standards board could also question how widely some of the more dubious techniques should be used. Mary Bush, a forensic dentist at the State University of New York in Buffalo, says that there is little evidence that bite marks left in skin can reliably identify perpetrators. In her lab, moulds of different sets of teeth were clamped into the skin of cadavers. Digital images of the marks were then analysed. Often, the marks could not be used to identify the teeth responsible.

Gregory Golden, president of the American Board of Forensic Odontology, argues that the method is useful for eliminating suspects or determining whether a bite mark is human.

According to the Innocence Project, however, at least 15 people whose convictions involved bite marks and who served time in prison have been exonerated through DNA evidence since 1993. That alone suggests that the method should be investigated, says Bush. "We're fighting 30 years of precedent." ■

DENNIS STONE/REX

Immigrants and visitors to the United Kingdom often face red tape and discouraging policies.

IMMIGRATION

UK visa problems worry scientists

Immigration policies scare off foreign talent, warn critics.

BY DANIEL CRESSEY

The United Kingdom's increasingly tough stance on immigration is driving foreign scientists to competing nations, the academic community has warned.

At a meeting with the Home Office last month, representatives of leading universities and scientific organizations said that unwelcoming government rhetoric about reducing immigration, together with complicated visa procedures for visiting researchers, make Britain an unattractive destination for scholars.

The Campaign for Science and Engineering (CaSE) in London, which promotes science-friendly policies and coordinated the meeting, is now actively lobbying the government to change its policies to avoid scaring away international students and academics. The House of Lords, the upper chamber of Parliament, has started an investigation.

"The really big issue is the one of how the UK is perceived internationally and how attractive it seems to people who wish to come here," says Sarah Main, director of CaSE.

She adds that CaSE's members, which include universities, scientific societies and businesses, have expressed concern both about the complexity and bureaucracy of certain visa schemes, and more generally "about the welcome being offered to often very senior academics and professionals" when they try to come to the United Kingdom.

In 2010, CaSE launched a campaign to keep the United Kingdom open to scientists (see *Nature* **468,** 346; 2010). Subsequently, the government altered various rules; for example, it exempted employers of PhD-level staff from a requirement to prefer candidates who already have UK residency.

"We're now viewed overseas as quite a potentially unwelcoming place to be."

In addition, a whole new visa type — the 'exceptional talent route' — was launched to attract skilled migrants (see *Nature* **476,** 243; 2011). This created up to 700 places per year for scientists to enter the United Kingdom, if they are endorsed by the Royal Society, the Royal Academy of Engineering or the British Academy. There are also 300 places for people working in the arts.

But the success of these moves has been

mixed, and in recent months the government has been taking a harder line on immigration. It now wants to reduce net migration to the United Kingdom from 182,000 people in 2012 to tens of thousands per year by 2015.

Susan Kay, executive director of the Engineering Professors' Council in Horsham, says that scientists complaining about the immigration system "have been getting louder". "We're now viewed overseas as quite a potentially unwelcoming place to be for academics," she says.

The visa system "needs to be simpler, it needs to be more accessible", says Kay, who was at the CaSE meeting and was "very much encouraged" by the Home Office's receptiveness. In a statement, immigration minister Mark Harper said that the government was building a system that "supports growth by curbing abuse, while still welcoming the brightest and the best".

However, the Home Office may argue that the scientific community has not exploited the concessions that it was granted in 2010. The exceptional-talent system has been hugely undersubscribed, with only 89 applicants in total entering the United Kingdom through this route between its launch in 2011 and June 2013. The reasons for this are unclear.

"It's a shame. It means universities are missing out on exceptionally talented people who could be using that route. Second, they've lost the argument with the Home Office if they don't make this work," says Ian Robinson, a senior manager at immigration law firm Fragomen in London. He helped to design the exceptional-talent route while working at the Home Office.

COMPLEX SYSTEM

Robinson says that the UK system has many advantages. It is based on checking applicant-supplied evidence against set criteria, so it provides certainty that can be lacking in more subjective methods, such as the US and Australian interview systems (although US immigration rules are currently being reformed; see 'Principles for compromise'). It is also fast — according to figures from Fragomen, a work-visa application to the United Kingdom is processed in an average of 15 days or fewer, versus 46–60 days for France and Germany, and 76 or more for Spain and Italy.

Overall, says Robinson, concerns are "largely down to perception. Generally the system does work. Myth-busting is needed — the Home Office and the scientific community need to go out there and say this is the system and this is how you use it."

Student immigration is also a source of concern, with statistics released last month showing that the number of students coming to the United Kingdom from outside the European Union fell from 302,680 in 2011–12 to 299,970 in 2012–13 — the first recorded drop ever. Rules on student visas have been toughened up in the past few years; for example, in 2011 restrictions were placed on graduates who stay on to work after they complete their studies.

A House of Lords committee is this week holding the first evidence session of an inquiry into whether visa problems are deterring science and engineering students. John Krebs, who heads the committee, told *Nature* that there is "ongoing concern" about the issue.

Another concern is the problems faced by scientists seeking to visit the country to attend conferences or give talks, who fall into the 'academic visitor' visa category. In January to September last year, the UK Border Agency received 4,770 applications for such visas, and rejected 625 (13%). The grounds for rejections are not published, but the rejection rate has stayed roughly consistent for the past six years.

Denis Noble, president of the International Union of Physiological Sciences and a systems biologist at the University of Oxford, was involved in organizing last year's union meeting in Birmingham, which attracted around 3,200 delegates. But in the weeks before the conference, he says, he spent nearly all of his time dealing with visa problems experienced by around 40 people who wanted to attend.

"From the experience I've had, there is an image problem in the United Kingdom. It's important that the right people try to put that right," he says. ■ SEE EDITORIAL P.6

Additional reporting by Richard Van Noorden

US IMMIGRATION REFORMS

Principles for compromise

In a move that may aid scientists who want to live and work in the United States, Republican party leaders in the House of Representatives released a set of principles for immigration reform on 30 January. It marks a tentative sign that they might consider compromising with broader efforts passed by the Democratic-controlled Senate.

Under current policy, the thousands of scientists and engineers who seek permanent residence in the United States must compete for a total of 140,000 'green cards' each year. Per-country limits on those permits often leave applicants from China, India and other oversubscribed countries waiting years for approval.

A Senate bill passed last June would relieve the green-card backlog by eliminating country-based caps (see *Nature* **499,** 17–18; 2013). The measure would also create an unlimited number of green cards for immigrants with master's or doctoral degrees in science, technology, engineering or mathematics (STEM) subjects from US universities.

But the House, controlled by Republicans, baulked at the sweeping plan, which included a controversial path to citizenship for illegal immigrants. It chose to pursue smaller immigration proposals. One measure, in the works since last year, agrees with scrapping country limits on green cards, but would give 55,000 spots to holders of advanced STEM degrees from US universities.

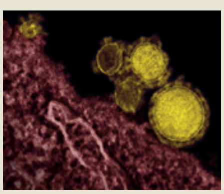
For any of the bills to become law, the two parties must overcome deep divisions on issues that have nothing to do with visas for scientists. The principles floated last week include options for granting limited legal status to illegal immigrants — a significant step towards the Democrats' position, and a sign that negotiations on immigration reform could restart.

But immigration reform has eluded lawmakers for years, and few observers are holding their breath. "It's hard for me to get too excited right now, until we start seeing people start to come out and say more," says Benjamin Corb, director of public affairs for the American Society for Biochemistry and Molecular Biology in Rockville, Maryland. **Helen Shen**

MORE ONLINE

TOP STORY

NIAID



First mouse model with coronavirus receptor go.nature.com/6osqq7

MORE NEWS

- Academics report reading fewer scholarly papers for first time in 35 years go.nature.com/nyrwus
- Metagenomics finds Beijing smog laced with microbes go.nature.com/eyilt8
- Antioxidants can promote tumour growth go.nature.com/suvue9

NATURE PODCAST

Sound tourism; the deceptively greener Amazon; and graphene superhighways nature.com/nature/podcast

Some contributors to citizen-science initiatives, such as Project Noah, sport a tattoo of the project's logo.

PROJECT NOAH/DAN DOUCETTE

CITIZEN SCIENCE

Computer sharing loses momentum

Competition and education needed to keep people engaged.

BY NICOLA JONES

The family of '@home' volunteer computing projects is growing ever more diverse. Spare time on a personal computer can now be donated to anything from finding alien life to crunching climate models or processing photos of asteroids. But enthusiasm is waning. The 47 projects hosted on BOINC, the most popular software system for @home efforts, have 245,000 active users among their 2.7 million registrants, down from a peak of about 350,000 active users in 2008 (see 'Slumping @home').

David Anderson, the founder of BOINC (Berkeley Open Infrastructure for Network Computing) and a computer scientist at the University of California, Berkeley, has several explanations for the slip. He says media coverage has declined now that volunteer computing is more than 15 years old. A shift to mobile-computing devices has probably also hurt — BOINC can run on an Android phone while charging, but uses too much battery power when unplugged. And the site has been unable to attract a broad demographic of volunteers.

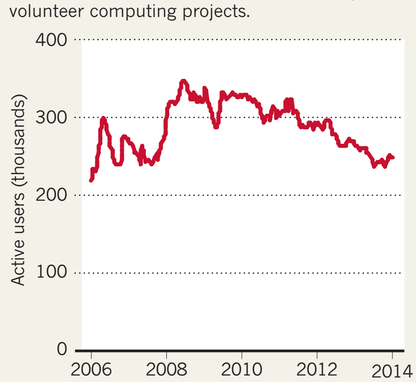
"Essentially, we have a bunch of middle-aged, male computer nerds," says Anderson. "We have thought long and hard about ways to break out of that category, using Facebook, for example, but none of that has been all that successful."

On 20–22 February, at the 3rd Citizen Cyberscience Summit in London, conference-goers will trade tips on how to entice volunteers into projects ranging from BOINC-style distributed computing to more-active 'citizen-science' projects, in which users are asked to donate not just their time but also their brains.

The desire to keep numbers up is not just academic. If distributed computing flourishes, serious money can be saved, says Francois Grey, coordinator of the Citizen Cyberscience Centre, based in Geneva, Switzerland. He notes that the Chinese Academy of Sciences in Beijing has been monitoring the economic benefits of CAS@home, which uses volunteers' computing time for projects such as predicting protein structures. The academy estimates that US$20 million has been saved since it launched CAS@home in September 2010, by using donated computing power rather than buying it from a company such as Amazon.

SLUMPING @HOME

The past several years have seen a decline in the number of active users in the BOINC family of volunteer computing projects.



WILLY DE ZUTTER, BOINCSTATS

Grey predicts that funding bodies might at some point enforce the use of volunteer computing whenever possible, rather than allowing grant money to be used for supercomputer time or cloud-based services. "It's very delicate. There are big IT companies with vested interests in selling supercomputers to universities," he says. "But I think it's something that will happen at some point."

For volunteer computing to be used in a bigger way, participation rates need to keep up. Perhaps the most obvious motivator — money — is deemed a bad idea. "Small amounts of money are too trivial, and may be almost insulting," says Grey. "It goes against the idea of volunteering." Only one BOINC project — IBM's World Community Grid, an umbrella initiative that oversees a batch of biomedical projects aimed at goals such as drug discovery — has partnered with a scheme that allows volunteers to earn virtual cash (which can be exchanged for real money) for their time. This had a measurable but small overall impact, says Anderson, earning the grid as many as 15,000 new volunteers, bringing the total so far up to almost 650,000.

A more powerful motivator is pleasure. This can be achieved by turning participation into a game. FoldIt, for example, asks volunteers to optimize protein folding, which requires a mix of intellect and intuition that some describe as similar to chess. Competition can also provide pleasure. Many projects offer scoreboards and awards such as virtual titles or badges to mark progress; some people have become so devoted that they have had the badges tattooed on their bodies. In the BOINC world, groups of volunteers have formed teams that compete to donate the most time over a designated period. These competitions offer a short-term boost, but the effect wears off, says Anderson.

Engaging participants in the core science mission is by far the best motivator, says Oded Nov, who studies links between new technologies and human behaviour at New York University. That includes giving participants credit in scientific papers and showing them how their help is advancing research. The World Community Grid, for example, hosts regular Q&A sessions with its project scientists. "Education is a great motivator," says Nov.

That could be one reason why the Zooniverse — the largest host of citizen-science schemes — has not seen a decline in participation. Its family

of 22 projects asks volunteers to do everything from identifying galaxy types in astronomical images to transcribing historical weather records. Robert Simpson, a developer and head of communications for the Zooniverse team, says that the five-year-old scheme has 930,000 registered participants and that there is fairly consistent interest in new projects.

Quantifying the effects of different motivational tools is difficult, says Grey, whose cyberscience centre has received funding to explore the possible benefits of common rules and credit schemes across different platforms such as BOINC and the Zooniverse. "Because of its grass-roots nature, everyone's doing their own thing; there's no common metric," he says.

One thing is certain: there is still plenty of spare brainpower to access. "US citizens alone spend 200 billion hours watching television a year," says Simpson. "We only need to tap a tiny fraction of that." ■

PUBLISHING

Elsevier opens its papers to text-mining

Researchers welcome easier access for harvesting content, but some spurn tight controls.

BY RICHARD VAN NOORDEN

Academics: prepare your computers for text-mining. Publishing giant Elsevier says that it has now made it easy for scientists to extract facts and data computationally from its more than 11 million online research papers. Other publishers are likely to follow suit this year, lowering barriers to the computer-based research technique. But some scientists object that even as publishers roll out improved technical infrastructure and allow greater access, they are exerting tight legal controls over the way text-mining is done.

A few years ago, scientists complained that publishers were stymieing ambitious plans to use computer software to pull out information from published papers. Some researchers who ran software to harvest data from online articles found their programs blocked, and those who asked for permission found themselves trapped in tortuous case-by-case negotiations — even though they had already paid subscription fees for access. Max Haeussler, a computational biologist at the University of California, Santa Cruz, for instance, spent more than three years arguing with publishers for permission to extract DNA data from 3 million articles to annotate an online map of the human genome (see *Nature* **483,** 134–135; 2012).

"It was a legitimate criticism, that people sent text-mining requests in to publishers and they bounced around for a time without any response," admits Chris Shillum, vice-president of product management for platform and content at Elsevier. The publisher previously considered requests "case by case", he says — but it now wants to make text-mining permissions quicker and easier to obtain. "What we've tried to do is take the practical barriers away."

Under the arrangements, announced on 26 January at the American Library Association conference in Philadelphia, Pennsylvania, researchers at academic institutions can use Elsevier's online interface (API) to batch-download documents in computer-readable XML format. Elsevier has chosen to provisionally limit researchers to 10,000 articles per week. These can be freely mined — so long as the researchers, or their institutions, sign a legal agreement. The deal includes conditions: for instance, that researchers may publish the products of their text-mining work only under a licence that restricts use to non-commercial purposes, can include only snippets (of up to 200 characters) of the original text, and must include links to original content.

> ***"Finally, someone is showing that there is no need to be afraid of text-mining analysis."***

"Finally, someone is showing that there is no need to be afraid of text-mining analysis any more," says Haeussler.

Researchers working on the Human Brain Project — a European consortium that plans to use a supercomputer to recreate everything known about the human brain — have already used Elsevier's interface to do text-mining, says the project's spokesman Richard Walker, who is based at the Swiss Federal Institute of Technology in Lausanne. "We are very pleased with it. It resolves genuine technical issues," he says.

And neuroscientist Shreejoy Tripathy at the University of British Columbia in Vancouver, Canada, worked with Elsevier last year to pull out information on neuron physiology from thousands of articles (see neuroelectro.org). Text-mining is not yet well known, he says, but he hopes that the easier access will kick off its greater adoption among scientists. "As more papers get published that use text-mining, other researchers like myself — who are neuroscientists and not programmers — will see the need for the technique," he says.

Shillum says that Elsevier is ahead of the curve — but that other publishers are likely to follow soon. CrossRef, a non-profit collaboration of thousands of scholarly publishers, will in the next few months launch a service that lets researchers agree to standard text-mining terms and conditions by clicking a button on a publisher's website, a 'one-click' solution similar to Elsevier's set-up.

And, in the past year, large institutions and pharmaceutical companies have started to ask for text- and data-mining rights when renegotiating site licences, says Jessica Rutt, rights and licensing manager at Nature Publishing Group (NPG), the publisher of this journal. Anyone with those rights may mine NPG content. Many publishers are also experimenting with delivering text-minable content to pharmaceutical companies for an extra fee, she adds.

But some researchers feel that a dangerous precedent is being set. They argue that publishers wrongly characterize text-mining as an activity that requires extra rights to be granted by licence from a copyright holder, and they feel that computational reading should require no more permission than human reading. "The right to read is the right to mine," says Ross Mounce of the University of Bath, UK, who is using content-mining to construct maps of species' evolutionary relationships.

National governments are also weighing in on the issue. The UK government aims this April to make text-mining for non-commercial purposes exempt from copyright, allowing academics to mine any content they have paid for. And the European Commission, worried that barriers to computational research could hinder scientific innovation, is also examining the issue. It has convened a group chaired by Ian Hargreaves, an intellectual-property specialist at Cardiff University, UK, who recommended the changes to UK law, to examine the economic impact of text- and data-mining for scientific research and barriers to its use. The panel will reach conclusions by the end of February.

"Our plan is just to wait for the copyright exemption to come into law in the United Kingdom so we can do our own content-mining our own way, on our own platform, with our own tools," says Mounce. "Our project plans to mine Elsevier's content, but we neither want nor need the restricted service they are announcing here." ■

CHRISTOPHE SIMON/AFP/GETTY

The Domaine de Vassal vine collection near Montpellier holds 2,300 different grape varieties.

VITICULTURE

Grapevine gene bank under threat

Scientists raise concerns about relocation of premier French research vineyard dubbed the 'Louvre of vines'.

BY DECLAN BUTLER

Uncertainty hangs over one of the world's largest and most important grapevine collections. The Domaine de Vassal vineyard, on France's Mediterranean coast, houses a vast sweep of grape biodiversity that is essential to research and winegrowers in France and around the world.

The 138-year-old collection, managed by the French National Institute for Agricultural Research (INRA), has been threatened with eviction, prompting a decision to relocate it.

That is raising concerns among scientists and winegrowers, because money to pay for the prospective move — costing an estimated €4 million (US$5.4 million) — has yet to be found. Even then, the sheer logistical complexity is such that relocation is likely to take years to complete, says INRA, and means that much of its research may be put on hold.

Dubbed the 'Louvre of grapevines' by the local press, the vineyard near Marseillan, southwest of Montpellier, contains thousands of unique grape varieties. As well as having a conservation role in preserving genetic diversity, the collection is used for research and for breeding qualities such as flavour, colour, adaptation to specific regions and pathogen resistance. Several hundred samples from the Domaine de Vassal are used annually, mainly by other French labs, but also internationally.

"The collection is of utmost value to the international grapevine genetics community," says Carole Meredith, an emeritus geneticist at the University of California, Davis. "Although many countries have established collections of their own heritage grape varieties, the Vassal collection is among the oldest and best curated."

Meredith notes that much of her own research would have been "impossible" without this "living library". Her lab's previous studies of the vineyard's specimens revealed Chardonnay's somewhat undistinguished heritage — one of its parent varieties is a noble Pinot, but the other is a Gouais, a grape long shunned as mediocre (J. Bowers *et al. Science* **285,** 1562–1565; 1999).

The collection was started in 1876 by French researchers in response to a pest outbreak that saw the near-destruction of Europe's vineyards. The outbreak was caused by accidental introduction of phylloxera — an aphid that infests roots and kills the vine.

The vineyard was initially located near Montpellier, but moved to the Domaine de Vassal in 1949, where it expanded greatly. It now houses some 7,500 accessions from 47 countries, representing 2,300 different grape varieties, including wild species, rootstocks, hybrids and mutants.

But negotiations with the vineyard's landowner, wine company Domaines Listel in Sète, near Montpellier, have broken down over the renewal of the 30-year lease on the 27-hectare site. In 2011, Domaines Listel issued an eviction notice; in 2012, INRA took the dispute to the agricultural land tribunal in Béziers, which is scheduled to hear the case this June.

Yves Barsalou, president of Domaines Listel, says that the company remains "open to all discussions" to find a solution that allows INRA to remain at the nearshore site.

In December, INRA announced its intention to relocate the collection, probably to a site alongside Pech Rouge, an INRA viticulture and oenology research station in Gruissan, about 70 kilometres southwest of Domaine de Vassal.

Olivier Le Gall, INRA's deputy director-general in charge of scientific affairs, says that the agency is "extremely committed" to preserving the collection, and is likely to have to find most of the moving costs itself. Other possible funding sources, he says, may include the French Vine and Wine Institute in Grau de Roi, which does applied viticulture and wine research.

The relocation, which should get under way this year, will be technically complex, says Jean-Michel Boursiquot, a vine taxonomist at Domaine de Vassal. Many specimens were collected as urgent rescue cases and carry diseases, but are protected from full-blown infections at the Domaine de Vassal because they are grown in beach sand. The sand shields against root infestations of phylloxera and nematode worms that can spread devastating viral vine diseases.

INRA has decided against a similar nearshore location for the vineyard, fearing that rising sea levels caused by climate change would make the site vulnerable to high salinity and flooding, says Boursiquot. At Pech Rouge, the plants will grow on higher ground in limestone soils. This will leave diseased plants susceptible to root infestations, so INRA intends to render the collection disease-free, a laborious process that involves repeated culturing and then propagating each plant until it is without pathogens. "It's an enormous job, which to our knowledge has never been done on such a scale," says Boursiquot. He thinks that this cleaning process — equating to half the move costs — will take 5–10 years.

Mark Thomas, a grapevine researcher at the Commonwealth Scientific and Industrial Research Organisation's Waite campus in Urrbrae, Australia, says that the Domaine de Vassal is one of the few grapevine germplasm collections to have been extensively characterized genetically, using DNA fingerprinting. This makes it an international reference source, and allows researchers to explore the genetic relationships between varieties, and their origins.

"This foundation of information is of great use for those around the world seeking to breed improved grape varieties," adds Bruce Reisch, who develops such new strains at Cornell University's research station in Geneva, New York. "It's extremely important that this collection be preserved well into the future." ■

SEMICONDUCTORS

Phosphorene excites materials scientists

Physicists look past graphene for atom-thick layers that could be switches in circuits.

BY EUGENIE SAMUEL REICH

Graphene, a one-atom-thick layer of carbon, has charmed materials scientists with its enticing electrical properties that allow electrons to flow freely across its surface. But the material lacks a natural band gap — a range of energy states in which electrons cannot exist freely — that could be used to switch this flow on and off. This reduces graphene's usefulness as a replacement for the semiconductor switches in computer circuits.

Last month, research groups in the United States and China reported[1,2] on work towards a promising candidate that could fulfil both needs: phosphorene, an atom-thick layer of the element phosphorus that does have a natural band gap. The work is part of a trend that David Tománek, a condensed-matter theorist at Michigan State University in East Lansing, dubs the "post-graphene age" — in which researchers are exploring alternatives in the hope of overcoming graphene's deficiencies. The rationale is that phosphorene might be useful for making thin, flexible electronics that could be more easily cooled than silicon ones.

"There is quite a lot of hype in this area."

Physicists have been studying black phosphorus — a layered material held together by weak chemical bonds — since the 1960s. But it was only last year that they began trying to isolate single layers. Just as in graphene, phosphorene atoms are arranged hexagonally, but in phosphorene the surface is slightly puckered. With its band gap, phosphorene can be switched between insulating and conducting states, and it is still flat enough to confine electrons so that charge flows quickly, leading to a relatively high mobility that is prized by electrical engineers.

Two groups, one[1] led by Peide Ye of Purdue University in West Lafayette, Indiana, and the other[2] by Yuanbo Zhang of Fudan University in Shanghai and Xian Hui Chen of the University of Science and Technology of China in Hefei, posted reports on a preprint server in January. They reported that they had stripped black phosphorus to two or three atomic layers by using sticky tape to peel the layers off a larger sample — the same method used in 2004 to isolate layers of graphene. Neither team has yet isolated a single layer of phosphorene.

There are reasons for optimism, however. Already, the groups have reported charge flows at speeds comparable with those in single layers of molybdenum disulphide, a semiconductor material with a band gap that has been tinkered with for nearly two decades. And phosphorene, unlike molybdenum disulphide, is made from a single element, so pure samples are, in theory, easier to obtain.

Phosphorene shares this purity with other post-graphene contenders such as silicene, made from silicon, and germanene, made from germanium. Although both of these are predicted to facilitate speedier charge flows than phosphorene, neither has a natural band gap. Both needs could be met by yet another material: stanene, a single layer of tin predicted by theorists[3] in 2013 that has not yet been created.

A problem for all of these materials is their instability, because single layers can react with the air. Silicene, a favourite among post-graphene researchers, was stabilized in 2012, but it is still hard to prevent electrical interference from the metallic substrates it has to be grown on, says Patrick Vogt, a physicist at the Technical University of Berlin — so the applications imagined for silicene are a way off. "There is quite a lot of hype in this area," he says.

Phosphorene seems more stable than its competitors, but it is not easy to produce: making black phosphorus entails putting the raw, powdered element under extreme pressure. Phaedon Avouris, a chemical physicist at IBM's Thomas J. Watson Research Center in Yorktown Heights, New York, says that the latest results justify more study, but he suspects that phosphorene's success in electronics will depend on whether researchers can find efficient ways to extract single layers and deposit them on substrates.

Sébastien Francoeur, a physicist at the Polytechnic Institute of Montreal in Canada, has already been seduced. He began working with black phosphorus after seeing the latest results. "A two-dimensional material that is a semiconductor is interesting technologically," he says. ■

1. Liu, H. *et al.* Preprint at http://arxiv.org/abs/1401.4133 (2014).
2. Li, L. *et al.* Preprint at http://arxiv.org/abs/1401.4117 (2014).
3. Xu, Y. *Phys. Rev. Lett.* **111,** 136804 (2013).

DOWNHILL FORECAST

WINTER SPORTS FACE AN UNCERTAIN FUTURE AS THE PLANET WARMS.

BY LAUREN MORELLO

Skiers, snowboarders and other athletes got a bit of a shock when they arrived in Sochi, Russia, for the 22nd Olympic Winter Games. On the way into town from the airport, competitors passed rows of palm trees that thrive in the breezes blowing off the Black Sea. Forty kilometres away, on the ski slopes of Rosa Khutor, Sochi organizers have spent a year stockpiling manufactured snow as a hedge against the region's mild climate.

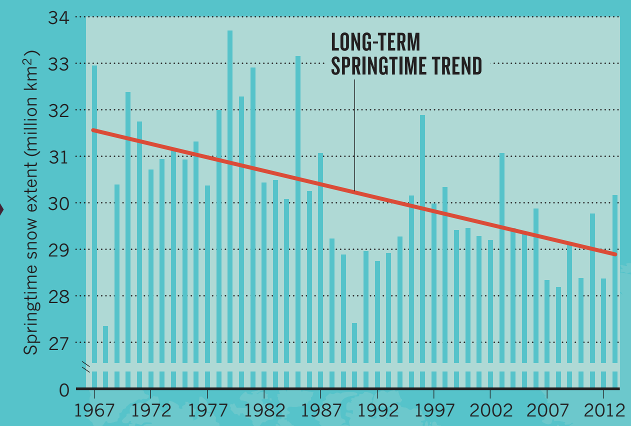
Meteorologists may scoff at the decision to hold a winter sporting event in a city where February averages a balmy 6 °C and temperatures hover just above freezing in the nearby mountains. But Sochi and its massive snow-making operation offer a glimpse of the future of skiing and the pressures that will confront Olympic planners as the world heats up.

"We know things are going to get warmer, and eventually, when you have temperatures above freezing more commonly than not, you're going to see less snow," says David Robinson, a hydroclimatologist who runs the Global Snow Lab at Rutgers University in Piscataway, New Jersey. In the Northern Hemisphere, the snow season has shrunk by about three weeks since the early 1970s (see 'Shorter winters'), and snow cover is

CONTINUED ON P.22

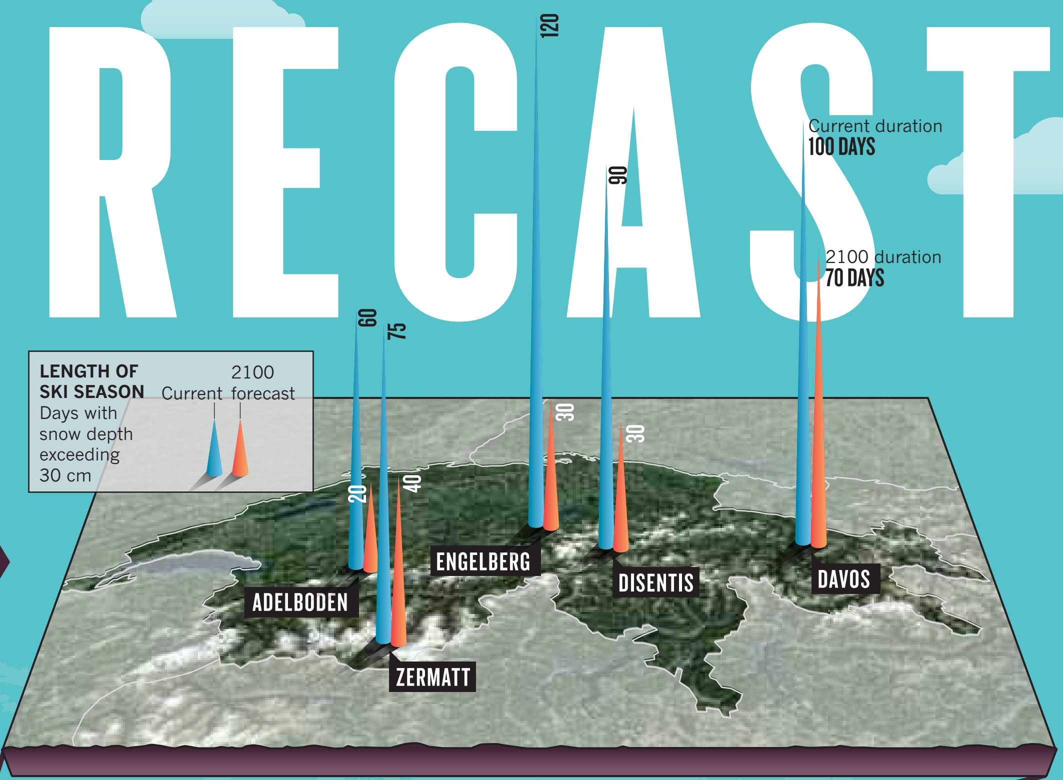
SHORTER WINTERS

Snow extent in the Northern Hemisphere has increased slightly in autumn and winter over recent decades, but has dropped substantially in spring, resulting in a shorter snow season.



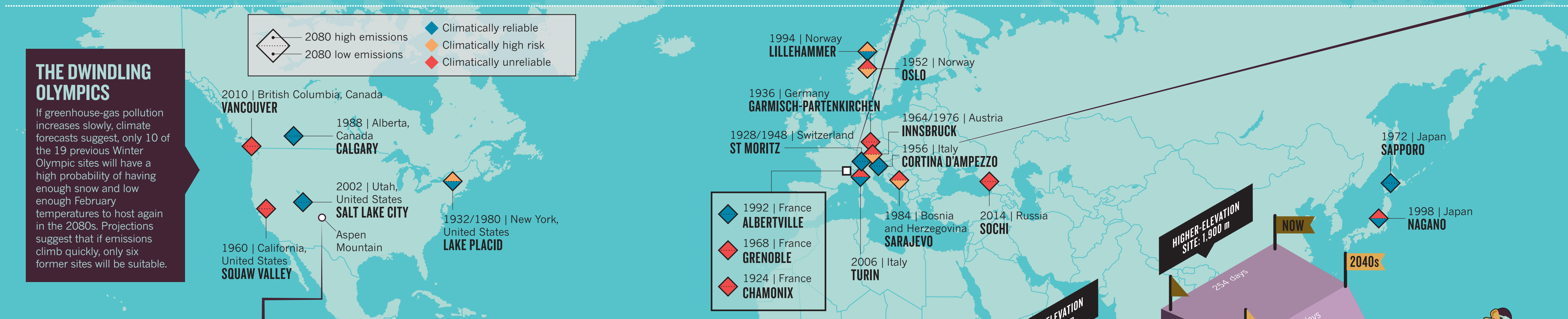
TROUBLE IN THE ALPS

If emissions of greenhouse gases continue to increase quickly, climate simulations project that many Swiss resorts will see a sharp reduction in the number of days with enough natural snow for skiing.



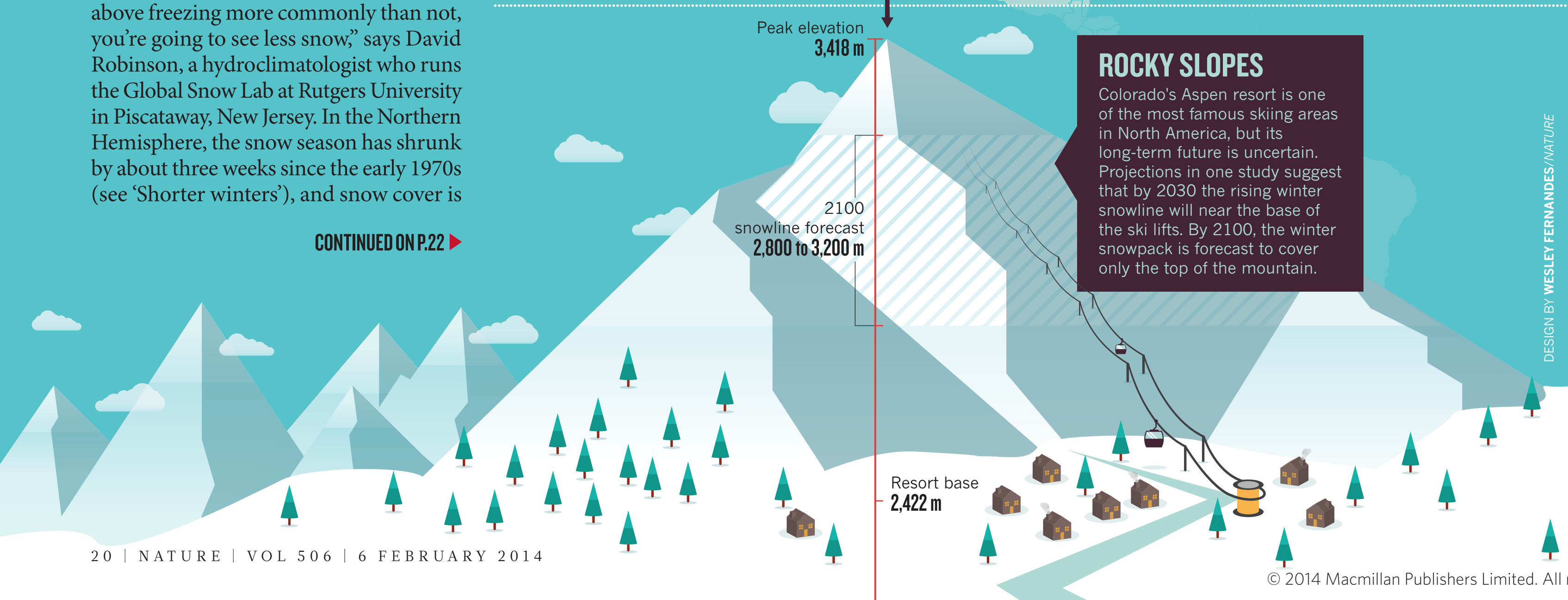
THE DWINDLING OLYMPICS

If greenhouse-gas pollution increases slowly, climate forecasts suggest, only 10 of the 19 previous Winter Olympic sites will have a high probability of having enough snow and low enough February temperatures to host again in the 2080s. Projections that suggest that if emissions climb quickly, only six former sites will be suitable.



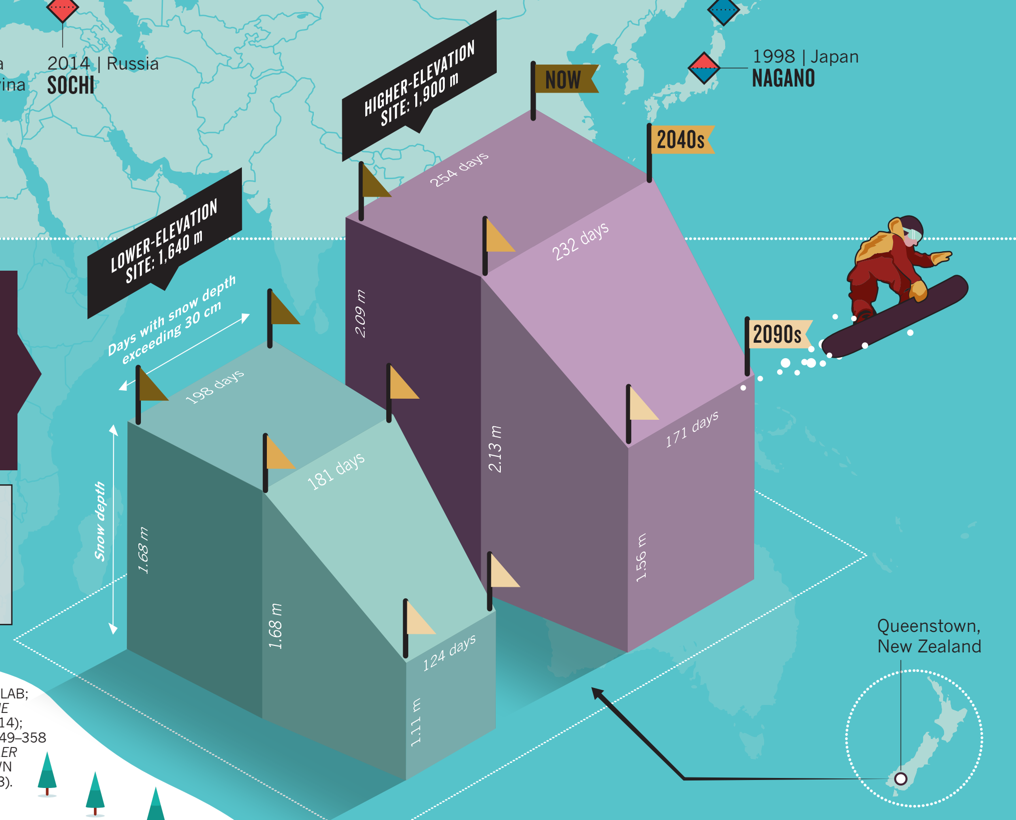
ROCKY SLOPES

Colorado's Aspen resort is one of the most famous skiing areas in North America, but its long-term future is uncertain. Projections in one study suggest that by 2030 the rising winter snowline will near the base of the ski lifts. By 2100, the winter snowpack is forecast to cover only the top of the mountain.



SNOW DOWN UNDER

Forecasts using a mid-range emissions scenario suggest that high-elevation skiing areas in New Zealand could see deeper snow in the 2040s. But by the 2090s, the number of snow days and the depth of the snowpack are projected to fall substantially.



DESIGN BY WESLEY FERNANDES/NATURE

SOURCES: SHORTER WINTERS: RUTGERS UNIV. GLOBAL SNOW LAB; THE DWINDLING OLYMPICS: D. SCOTT *ET AL. THE FUTURE OF THE WINTER OLYMPICS IN A WARMING WORLD* (UNIV. WATERLOO, 2014); TROUBLE IN THE ALPS: M. BENISTON *WIRES CLIM. CHANGE* **3,** 349–358 (2012); ROCKY SLOPES: B. LAZAR & M. WILLIAMS *PROC. WHISTLER 2008 INT. SNOW SCI. WORKSHOP* 998–1005 (2008); SNOW DOWN UNDER: J. HENDRIKX *ET AL. CLIM. CHANGE.* **119,** 965–978 (2013).

Snow has been stockpiled for the Olympic Winter Games in Sochi, Russia, for a year.

projected to decline substantially by the end of the century, according to a report released in September by the Intergovernmental Panel on Climate Change.

There is considerable uncertainty in regional forecasts for individual skiing regions, but climate researchers say that models agree on broad patterns. "By the second half of the twenty-first century, the models suggest we'll be seeing really big changes in temperature and precipitation," says Martin Beniston, a climate physicist at the University of Geneva in Switzerland. Already, signs of an unwelcome thaw have appeared at even the highest elevations. This season, the Verbier 4 Vallées resort in Switzerland eliminated two chair lifts after the lower edge of Tortin Glacier, at 2,800 metres elevation, receded by 40 metres in just 15 years.

The outlook is not so gloomy everywhere, at least initially (see 'Snow down under'). A warmer atmosphere can hold more moisture, so rising temperatures may actually increase snow at some high-elevation sites — such as the peaks of New Zealand and parts of the Swiss Alps — for several decades, until winter temperatures inch above freezing (see 'Trouble in the Alps'). In fact, average snowfall over the past decade at the Verbier resort has outpaced that in each of the previous three decades. But the area has also had to deal with more variability, warmer summers and a spate of hit-or-miss winters. "The extremes are much higher or much deeper," says Eric Balet, chief executive of Téléverbier, the company that owns 4 Vallées. That introduces an unwelcome element of unpredictability for resort managers.

Skiing areas at low elevations face the worst forecasts. The US states of Connecticut and Massachusetts are home to a combined 17 skiing areas, and a study suggests that by 2039, none will sustain a viable skiing season — defined by industry as 100 days or more — even with artificial snow-making (J. Dawson and D. Scott *Tourism Mgmt* **35,** 244–254; 2013). But at least 94% of the 18 resorts in the more northerly state of Vermont are projected to be viable until 2070 or beyond. A big difference is altitude: all the resorts in Connecticut and Massachusetts have a peak elevation below 750 metres, whereas 16 of those in Vermont exceed that, many by hundreds of metres.

> "THESE JUST AREN'T THE KIND OF CONDITIONS THAT PEOPLE GO SKIING FOR."

QUALITY AND QUANTITY

The prospect of losing small skiing areas worries Auden Schendler, vice-president for sustainability at Aspen Skiing Company, which runs the Aspen and Snowmass resorts high in central Colorado (see 'Rocky slopes'). "We call those feeder resorts," he says, because their lower prices and gentler slopes attract new skiers to the sport. "It doesn't serve us for other resorts to go out of business." And small mountains can produce big stars. The 2010 Olympic downhill champion, Lindsey Vonn, started skiing at Buck Hill, a 364-metre mountain in Burnsville, Minnesota; reigning World Cup slalom champion Mikaela Shiffrin trained as a child at New Hampshire's Storrs Hill Ski Area, where the peak elevation is just 177 metres.

It is not only the quantity of snow that is crucial for winter sports such as skiing and snowboarding. Quality matters too, says Anne Nolin, a snow hydrologist at Oregon State University in Corvallis. A warmer, moister atmosphere will produce heavier, wetter snow, not the dry, fluffy 'champagne powder' prized by many recreational skiers. Artificial snow created with snow-making cannons is often icy, perfect for laying the base of lightning-fast competition runs but less favourable for the average skier. And temperatures that skirt the freezing mark increase the risk that precipitation will fall as rain, not snow, and will raise the density of the snowpack. "These just aren't the kind of conditions that people go skiing for," says Nolin.

It is not clear what the changing face of winter portends for future Olympic Games (see 'The dwindling Olympics'). The competition has already been shaped by the vagaries of weather and climate, beginning with the decision decades ago to move figure skating, speed skating, ice hockey and curling indoors, says Daniel Scott, a geographer at the University of Waterloo in Canada. His research suggests that the pool of locations capable of hosting the games will shrink as the climate warms — and the colder mountain cities that may be the best fit may not have the infrastructure to handle a massive influx of athletes, spectators and organizers. That will force some difficult decisions, he says. "It's an interesting dilemma the International Olympic Committee will be caught in." ■

Lauren Morello *is the assistant US News editor at* Nature.

ILLUSTRATION BY GARY NEILL

THE CHANGING FACE OF PRIMATE RESEARCH

A hard-won political victory for primate research is at risk of unravelling in pockets of Europe.

BY ALISON ABBOTT

The worst moment in neuroscientist Andreas Kreiter's 16-year struggle to defend his research came when his wife arrived home after the birth of their second child. Waiting for her was an envelope containing a death threat against their three-year-old.

Kreiter, who uses macaques in his studies of the brain at the University of Bremen in Germany, is a veteran of the fierce and periodically violent tactics of animal-rights activists. When protests peaked in the late 1990s, he lived under police protection — but he still continued his research. "I had thought very carefully before deciding to work with primates," he says. "And I believe it is necessary if we are to understand the human brain."

Later Kreiter found himself facing an unfamiliar foe: local authorities looking to restrict primate research in their city. In 2008, Bremen officials declined to renew Kreiter's licence to work with macaques. The fate of his research has been in legal limbo ever since.

Kreiter's courtroom conflicts put him in good company. Across Europe, a particularly volatile patchwork of emerging local regulations threatens to distort the spirit of a recent European Union (EU) directive that explicitly allows research on non-human primates. Although some researchers say they have never felt so secure, others are facing new obstacles as activists change tack, from bullying researchers to putting pressure on regional policy-makers.

The problems continue even as the EU is pushing for the translation of basic research into therapies — a transition that often requires the testing of experimental therapies in primates. And opportunities for translational research are growing thanks to recent technological breakthroughs. However, restrictions on primate experiments could hinder their development.

Some European researchers are shifting their strategies, too, by talking more openly about their work with primates. But other scientists have simply stopped using monkeys altogether — or side-stepped the European quagmire by setting up controversial collaborations in other countries, particularly in Asia.

"Primate researchers should always expect to be under pressure, because we are handling a valuable and sensitive resource," says Roger Lemon of University College London, UK, who hopes his work on how the brain controls fine hand movements might lead to therapies for recovering function after a stroke. "But it's a sad irony that key developments may be transferring to countries that don't have the high level of animal welfare we have here."

STABILIZING STEP

The pressures on primate researchers have taken many forms. In the United States, for example, commercial airlines have effectively ceased all primate shipments by air within the country, making it difficult for researchers to transport animals. Many airlines in Europe have taken similar steps, but Air France continues to provide service.

Not long ago, the EU seemed to take a step towards stabilizing the environment for primate research. In September 2010, after more than a decade of anguished public debate, the EU adopted a directive governing the use of animals for research purposes. With its careful balance of animal-welfare and research needs, the directive seemed destined to ease tensions. Among other things, it established minimum welfare requirements for all animals, laid out definitions of pain intensity, and banned most research on great apes. It also included a hard-won clause — added at the last minute after intense lobbying by the biomedical community — explicitly permitting basic research on non-human primates, provided the work could not be carried out in any other species.

EU member states were required to anchor the directive into national legislation by 1 January 2013. And they were forbidden to 'gold-plate' the regulation by making national law stricter than EU law.

But animal-rights activists have continued their fight. They have honed their activities for greater media attention and have delayed implementation of the directive in several countries. Animal-rights organizations now focus on policy-makers rather than scientists, says Robert Molenaar, campaign manager for the Coalition Against Animal Experiments (ADC), which operates in the Netherlands and Belgium. The ADC is concentrating first on monkey research in universities, he says, because it is an easy way to get press coverage and influence political opinion.

The ADC is also forging international links and works closely with a sister organization in the United Kingdom, the Anti Vivisection Coalition (AVC), headed by Luke Steele. Steele spent nine months in prison after being convicted in 2012 of harassing staff at Harlan Laboratories, a contract research company in Blackthorn, UK. The jail time was interesting, he says: he used it to reflect on strategies. "Researchers themselves tend to be traditionalists who are not open to alternatives," he says. "I realised we need to go for policy-makers."

The AVC and the ADC were the main drivers of the Stop Vivisection Initiative, a petition calling for the EU animal-research directive to be abrogated and animal research to be banned altogether. The petition, launched in November 2012, collected more than a million signatures across the EU within a year. The signatures are now being verified; if they pass, the initiative will be granted hearings at the European Commission and the European Parliament.

> "YOU CAN'T GO DIRECTLY FROM MICE TO HUMANS. MICE ARE SIMPLY NOT A GOOD MODEL OF HOW PEOPLE SEE."

"This will reopen the debate — something we'd all rather do without, given the enormous effort that the commission, scientists and animal-welfare groups invested in achieving the compromise," says Stefan Treue, director of the German Primate Center in Göttingen and an adviser to the European Commission on the 2010 directive.

Treue doubts that the Stop Vivisection campaign will change European legislation — political demand for new therapies is too strong, he says. But, like many of his colleagues, he says that researchers working with monkeys should abandon their conventional tactic of keeping quiet, which cedes ground to the activists. Two months after the directive was approved, Treue helped to launch the Basel Declaration (see *Nature* **468,** 742; 2010), which commits its signatories — so far more than 2,500 — to be open about their animal research and to engage in public dialogue.

The declaration prompted a sea change, and many initiatives are emerging in its wake. For example, the Swiss Primate Competence Center for Research was launched last year in Fribourg to provide a training centre for scientists and technicians wanting to work with primates, and an educational one-stop shop for the public.

Individual scientists are also speaking up on their websites. Neuroscientist Pieter Roelfsema at the Netherlands Institute for Neuroscience in Amsterdam, who works with monkeys, says that so far activists have not targeted researchers in his lab. But he fears this may soon change. Last spring, minority parties in the Dutch parliament — including the Dutch Party for the Animals — posed formal questions about whether research using monkeys is necessary, if it could be replaced by alternative methods, and if the number of government-funded research institutes using monkeys could be reduced.

With these developments in mind, Roelfsema is planning a public-information webpage about the value of primate research, modelled on that of Nikos Logothetis, a director at the Max Planck Institute for Biological Cybernetics in Tübingen, Germany. Logothetis's site, which has thousands of visitors a week, emerged from a public-relations debacle. In 2009, he invited a team of investigative journalists from a national television company into his lab, imagining that the reporters would be impressed by his monkeys' luxurious accommodation, and surprised by how relaxed and content the animals seemed. Instead, the journalists portrayed a slightly mad scientist among suffering animals. The experience "spectacularly demonstrated the need for a reaction of scientific organizations to the escalating absurdity of the anti-vivisectionists", Logothetis says.

However, Tübingen — unlike Kreiter's Bremen — is a city where researchers enjoy a supportive political environment. Even the city's mayor, a member of the Green Party, which is not known for supporting animal experiments, has openly criticized flyers distributed by activists as untruthful, and described the harsh treatment of Logothetis as "unacceptable".

"This shows the power of local politics to influence how easy or difficult it can be to carry out research using monkeys in different European regions," says Treue, whose research centre also benefits from local political support in Göttingen. For scientists such as Treue, the EU directive has brought a feeling of stability.

THE ITALIAN JOB

That feeling is largely absent in Italy. In 2012, activists attacked a beagle-breeding facility near Brescia. It was later closed down. In 2013, they sabotaged experiments at the University of Milan. And last month, activists posted flyers that included photographs, addresses and phone numbers of some of the university's researchers in their home neighbourhoods.

By 2012, some populist politicians had adopted the animal-rights cause and used it to influence the Italian implementation of the EU directive. The proposed law went beyond the directive, calling for a ban on xenotransplantation and the use of animals in addiction research.

Italian scientists woke up late to the threat, and by the time researchers had organized a petition defending animal research — signed by 13,000 people in just a few weeks — the course of the distinctly gold-plated law was already set. It passed through parliament in December.

Researchers who use monkeys are also worried about ambiguities in how the Italian law

ANDREA RONCHINI/DEMOTIX/PRESS ASSOCIATION

Animal-rights campaigners have switched from targeting scientists to putting pressure on policy-makers.

interprets the EU directive's clause allowing research on non-human primates. "It's not clear at all whether basic research is allowed or not," says neurophysiologist Roberto Caminiti at the University of Rome La Sapienza, who chairs the Committee on Animals in Research for the Federation of European Neuroscience Societies.

The law also requires all research proposals involving non-human primates, cats or dogs to be authorized by the High Health Council (Consiglio Superiore di Sanità), the broad mandate of which includes drug licensing and approval of clinical protocols. This additional level of control, on top of the approval required from local ethical committees, would slow and destabilize the process, says Caminiti.

The legislation is expected to become law in March. As soon as it does, Caminiti and his colleagues plan to file an appeal to the EU Court of Justice. "Gold-plating is not allowed," he says, "so we are confident of winning." In the meantime, Caminiti predicts that Italian labs working with primates will all be able to argue that their work has health benefits for humans.

In Belgium, the government is hurrying through a similar gold-plating decree that would also ban the use of primates in addiction studies, and require a national committee to approve projects involving non-human primates, even after approval by local ethics committees. The Belgian health minister would have the final say on whether a particular project could go ahead, raising concerns that final decisions would be based on politics, rather than on science or ethics.

Political decisions are already affecting research in Switzerland, a non-EU country that is not bound by the 2010 animal-rights directive. In 2000, Switzerland's constitution was changed to protect the dignity of animals — a move that led courts to limit the use of monkeys to translational research.

Researchers in Fribourg have been able to continue their studies of spinal-cord repair in primates, but local authorities in Zurich have not renewed licences for basic research using primates since 2004. Kevan Martin, a director at the city's Institute of Neuroinformatics, had to stop mapping the functional microcircuitry of the macaque brain in 2006, when his licence expired. Martin was shocked to learn that local authorities had declined to renew his licence because the work was unlikely to reap practical benefits for society in the near term. He was even more shocked when his appeal to Switzerland's supreme court was turned down. "Is any applied research possible without basic research?" he muses.

WORKING ABROAD

In this climate, some Swiss scientists are relying on their collaborations in other countries to carry out primate experiments. Botond Roska of the Friedrich Miescher Institute for Biomedical Research in Basel and his colleagues have used mice to develop an experimental treatment for a common type of blindness called retinitis pigmentosa. The method is now poised for human trials, to be run by the small Paris-based biomedical company GenSight Biologics, which Roska co-founded. "But you can't go directly from mice to humans because you can't be sure if the neural circuits are the same," says Roska. "Mice are simply not a good model of how people see."

Rather than face uncertainty in Switzerland, Roska and his collaborators — GenSight and the Vision Institute in Paris — are conducting primate studies in France, where animal activists have less political support. Roska hopes the first human patient could be treated within the year.

Like Roska, Per-Olof Berggren at the Karolinksa Institute in Stockholm has reached a translational turning point in his research. He has developed an experimental therapy for diabetes in mice, and now needs to test it in primates before moving to humans. He thinks he could have got a licence for this in Sweden, but knew that he could not have afforded it. Regulations in the country, where animal-rights and animal-welfare groups are very powerful, require particularly large, sophisticated — and consequently expensive — primate facilities. So Berggren decided to do the work in Singapore, where he says facilities are first-class and ethical standards are as high as in Europe. "They have a long tradition of working with monkeys there, and it doesn't cost so very much."

Berggren is far from alone: many European researchers are taking their primate research to Asia, sparking a controversy that is dividing the scientific community. Some worry that standards of ethical oversight and animal welfare could be lower in certain Asian countries. And Martin points out that the trend exacerbates the loss of skills already apparent as the number of groups working on primates in Europe falls. (The number of primates used in the EU for scientific purposes shrank by more than 25% between 2008 and 2011, according to the European Commission.) "The loss is going to be much harder to reverse," he says. "Finding anaesthetists and surgeons has already become more difficult."

One European scientist, recently returned from two weeks at a leading institute in China, says that he found many Europeans setting up collaborations there — but they, like him, did not want to say so openly, for fear of damaging the reputations of their home institutions.

The scientist insists that ethical concerns are out of place, and that standards at the institutes match those of Germany and the United States. "It is not a question of low standards but of forward-looking research," he says. "And it is nice to enjoy the energy and optimism, and not always hear the word 'no'."

Back in Bremen, Kreiter still hopes to hear a 'yes' in court. With the support — moral and financial — of his university, he has spent more than five years fighting local authorities in a string of courtroom battles. He is now awaiting yet another verdict from a high court in Leipzig. "It may be the last," he says. "But you never know how things will develop." ■ SEE EDITORIAL P.5

Alison Abbott *is* Nature*'s senior European correspondent.*

COMMENT

ODD ANDERSEN/AFP/GETTY

A ship and containers washed ashore by Typhoon Haiyan, which ravaged the Philippines in November 2013.

Make supply chains climate-smart

Society's infrastructure is hit hard by extreme weather. Networks of trade, transport and production need to adapt globally, says **Anders Levermann**.

Extreme weather — including massive storms such as Typhoon Haiyan and Hurricane Sandy, and severe floods and droughts — is likely to become more frequent and intense as global warming accelerates[1].

Links in global economic chains and world markets mean that extreme weather in one place can have repercussions elsewhere. For example, a combination of exceptional rainfall and Cyclone Yasi in 2010–11 paralysed the world's fourth-largest region of coal exploration in Queensland, Australia. Coking coal prices rose the following year by 25%. In 2011, droughts and floods in Russia, Pakistan and Australia caused global food prices to climb, possibly contributing to the escalation of civil unrest in Egypt, Syria and Saudi Arabia. The long-term economic impacts of Typhoon Haiyan, which devastated the Philippines in November 2013, are yet to be felt, but are likely to affect global trade and manufacturing.

Yet the impacts of adverse weather on supply chains are missing from the assessments of the Intergovernmental Panel on Climate Change[1], and, with a few exceptions[2], ▶

are being ignored in discussions around adaptation. This is a mistake. Adaptation requires a global strategy, not just local ones.

It is these unanticipated and sudden shocks from extreme weather events on global trade that are most disruptive for society; gradual changes can be foreseen and are easier to adapt to. Sitting in a bathtub with the tap running it is easy to stop the floor getting wet as the water rises by placing a few towels (up to a point). But the effect of climate change is like throwing rocks into the water. Our interlinked societies, the dynamics of which we are only beginning to understand, are like dominos lined up on the edge of the tub. One wave can make them all tumble.

As protests in Brazil, Turkey and Greece in recent years have shown, societies do not have to be brought to the verge of starvation to descend into turmoil. Communities respond to events in unforeseen ways. The triggers might be small and the reactions cannot be understood with equilibrium economic theory[3]. With the fragility of our globalized economy becoming more evident, it is time to readjust our focus.

The influence of climate change on the worldwide flows of materials, electricity, communications and energy, including interactions between them and the rapid dynamics of volatile markets, needs to be modelled and understood. As a first step, we must collect and share basic data on global supply chains.

To this end, my colleagues and I at the Potsdam Institute for Climate Impact Research in Germany have set up a website called Zeean (www.zeean.net) to host such information and to help to kick-start a community effort to understand and model it. By making these economic data available, governments and companies will be alerted to crucial bottlenecks in supply chains and can respond accordingly. Factoring in the costs of climate change will, we anticipate, allow market forces to help to stabilize the global supply network and make societies more resilient.

SUDDEN SHOCKS

Almost all adaptation research so far focuses on local or regional responses to gradual changes in climate. This perspective has steered climate-change concerns towards long-term threats to fragile ecosystems and poor rural communities, which could be lessened through targeted strategies, such as altering crop production in the face of shifting monsoons in India, for example. Meanwhile, global trade connections and dramatic weather events are colliding — with even more costly consequences.

Thailand's devastating 2011 floods destroyed the country's automobile industry. And by disrupting manufacturing of mainly Japanese technology companies based in Thailand, the floods also caused a global shortage of hard-disk drives. With more than US$46 billion in damage[4], the floods were ranked as the fourth-costliest disaster ever by the World Bank. But that does not include secondary losses from missing means of production.

Disruption to pharmaceutical supply networks is already having deadly consequences. The increasingly complex supply chains for drugs are highly susceptible to blockages, causing shortages of medicines. In 2011, global supplies of cancer and AIDS medications ran short after a failed hygiene inspection of one US manufacturer (see go.nature.com/qhgfg5). Extreme weather events compound those risks. "Resting on old standards — even ones that have worked for decades — is no longer enough," cautioned Robert Parkinson, president of the US pharmaceutical company Baxter, addressing a US congressional subcommittee in 2008.

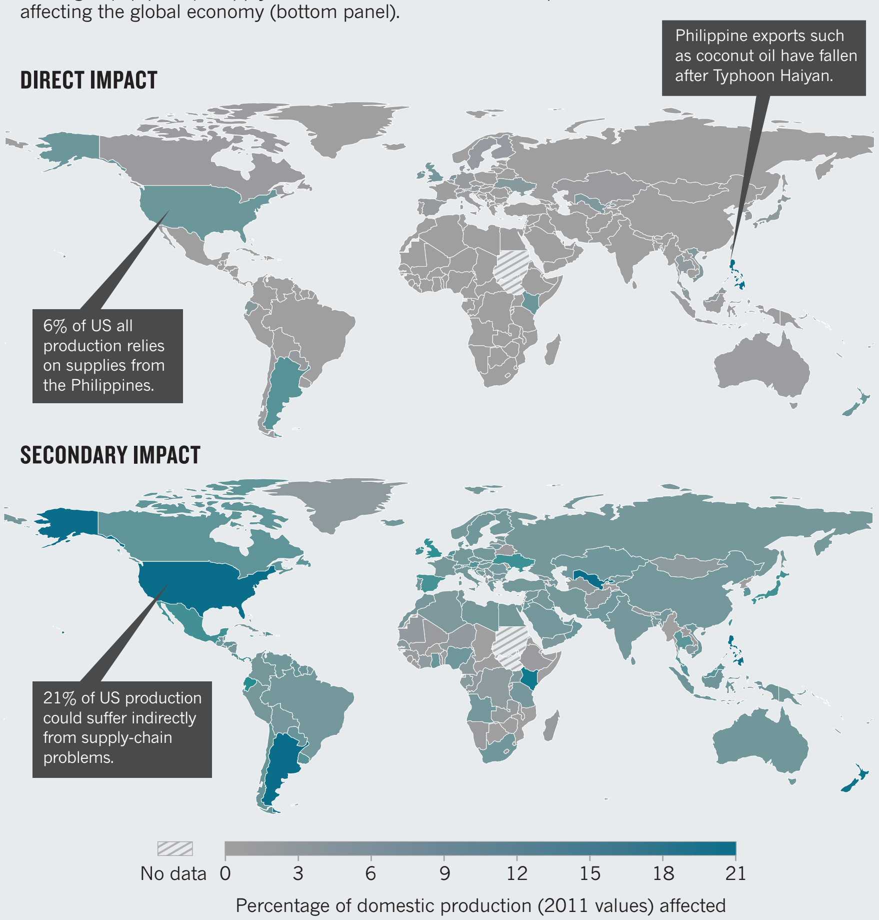
Bouts of severe weather in quick succession are even harder to recover from. Pakistan, for example, is still suffering from devastating monsoon-induced floods in 2010 and 2011. If hurricanes Sandy and Katrina had hit the US seaboard in the same season as last year's drought, even the United States might have struggled to cope.

KNOCK-ON EFFECTS

A simple calculation of economic flow disruption illustrates how the global repercussions of weather disruptions to supply chains are likely to eventually exceed the direct damages. On the basis of data collected by Manfred Lenzen, a professor of sustainability research at the University of Sydney, Australia, and his colleagues[5], we estimate that if not replaced, the cessation of exports from the Philippines, for example from fisheries and agriculture, would affect 6% of US production directly (see 'Global adaptive pressures'). The potential secondary effect, mainly through the retail trade, would be larger and could affect 21% of US production. The Philippines are globally the largest exporter of coconut oil, which is used in food products worldwide. For major economies such as Japan, Spain, the United Kingdom and the United States, mounting impacts on sectors such

GLOBAL ADAPTIVE PRESSURES

Simple modelling of supply chains shows how the cessation of exports from one country, for example the Philippines in the wake of a typhoon such as Haiyan, will affect many others. Direct trade links are broken immediately and may cause shortages (top panel). Supply restrictions from those nations spread further, affecting the global economy (bottom panel).



SOURCE: REF. 5

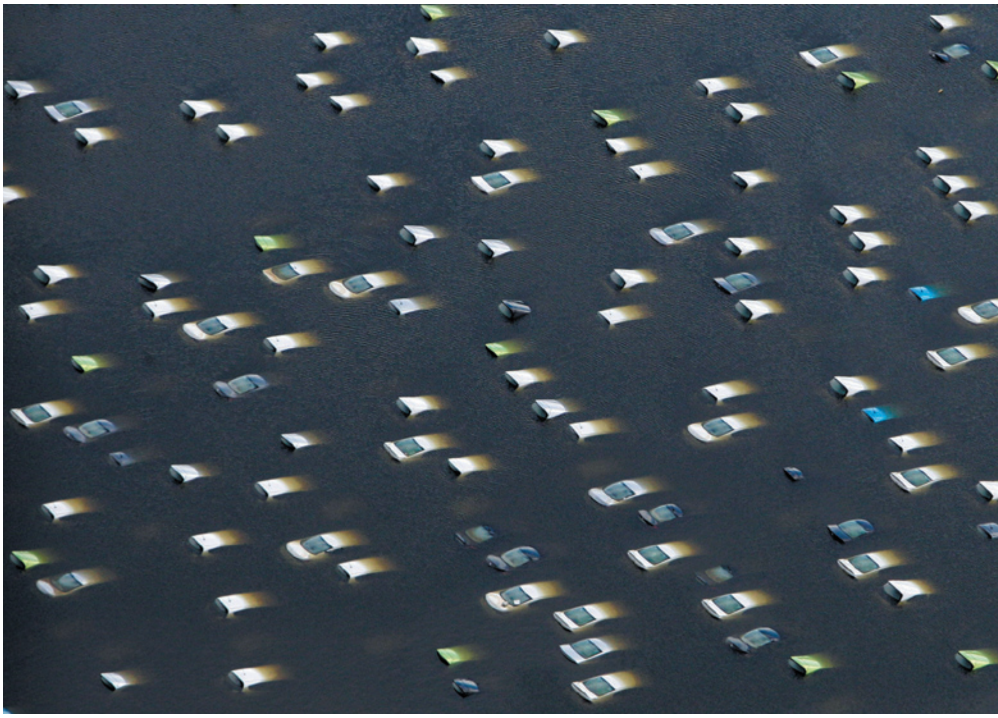
SAKCHAI LALIT/AP

as food production might greatly exceed the direct one.

Adaptation, like mitigation, needs to be global, but that does not mean it must be coordinated or regulated. Because the main aim of adaptation is to avoid damage, market forces might be effective in transferring costs. Just as the costs of insurance and meeting local safety requirements are factored into the prices of goods today, the costs of protecting supply chains from natural hazards can be included — if the information is available.

In many cases it will turn out that a diverse but reliable supply net is more favourable than relying on one source, even if that transportation route seems cheaper at first sight.

Lenzen's group has pioneered the application of supply-chain data to climate and sustainability problems[5]. They have compiled a database covering 26 industrial sectors in more than 180 countries, including, for example, the quantity of Australian coal used by the German steel industry and how much German steel is used in Russian ship building and Japanese car fabrication.



Submerged cars from a Honda factory after floods in Thailand in 2011.

COLLECT DATA

Building on Lenzen's data, we plan for Zeean to eventually cover 400 sectors and individual states, provinces and cities, allowing users to track the flows of specific goods at a scale appropriate for the effects of natural disasters. Users could ask, for example, how many batteries are shipped from Osaka, Japan, to California. Or what is the impact of a hurricane in Boston, Massachusetts, or a flood in Bangalore, India, on particular industries worldwide?

"With the wrong focus, we will protect the wrong places with the wrong tools."

In some cases, this supply-chain information is publicly available, but scattered around. In other cases, it must be deduced from import and export figures provided by national statistical agencies or government and industry bodies.

We intend the information posted to Zeean to be cross-checked and validated by registered and vetted users who will be assessed according to the quality of their input, in a similar way to websites such as Wikipedia. Single pieces of data, such as the number of cars produced in a region of Germany, for example, may be entered, as well as whole data sets, such as the 400-sector trade input–output matrices for Australia and Japan.

The information will need to be checked by comparing data from different sources or through consistency calculations. The sum of all individual flows out of a region should not exceed its total export, for example. Unlike in earlier approaches, we will not harmonize the information within a single global matrix, but will use the existing international data structures that comprise large economic sectors, and apply consistency tests to finer data where we can. This will avoid the introduction of artefacts, such as small unrealistic flows, to bridge data gaps and make computations feasible.

The processed information will be publicly available, so that small businesses and poor countries can use it. Only openly available data will be included, to avoid legal and rights issues. Open-source algorithms and analysis tools will be used for accessibility. Funding to maintain the database is being sought.

UNDERSTANDING RISK

With each piece of information added, Zeean will improve in accuracy. Users will be able to see a network evolving, analyse its connectivity and identify fragile links or nodes. By accommodating a variety of data, which need not be homogeneous, the database will allow for regional and economic foci of special interest.

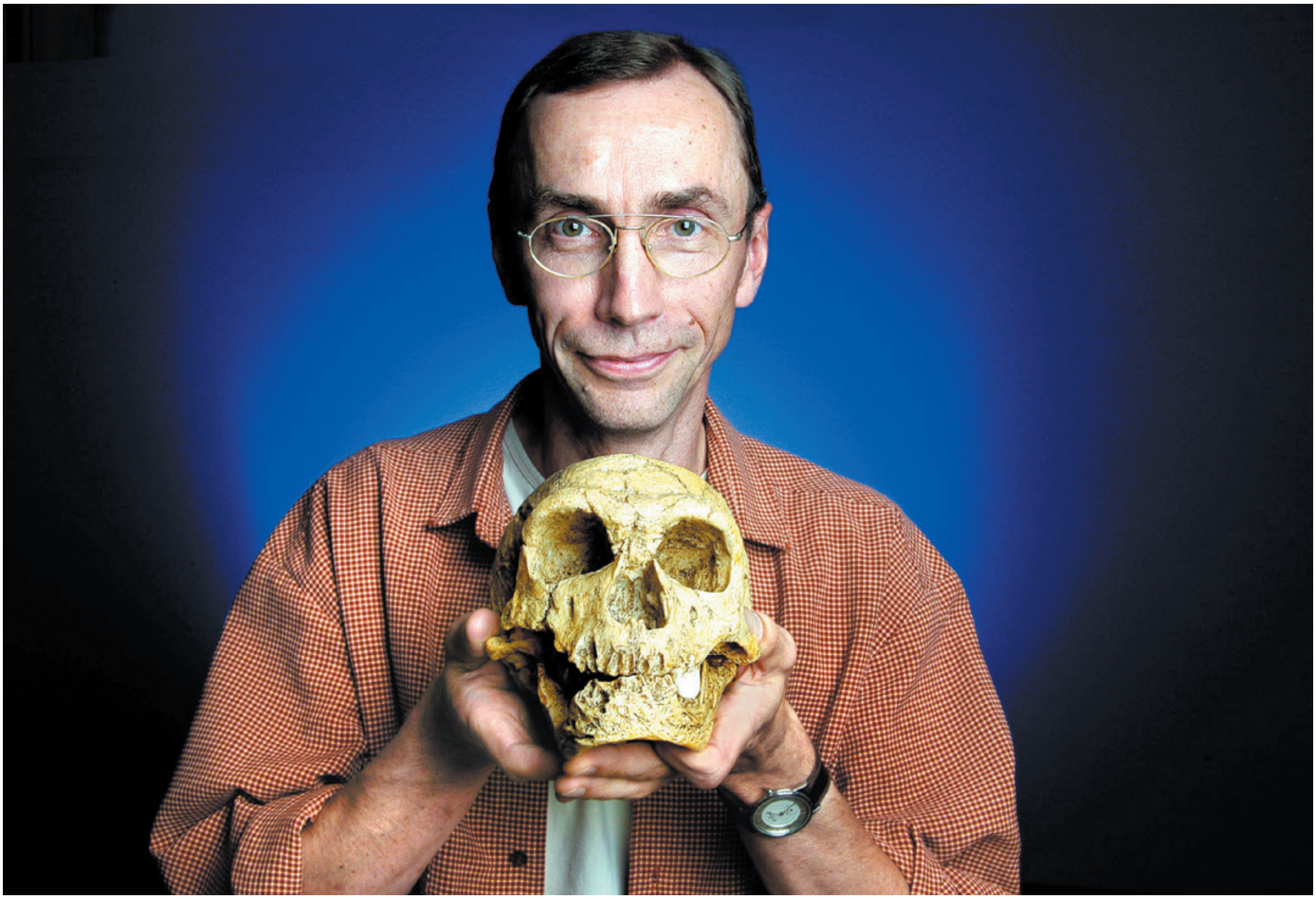
Longer term, to produce supply-chain risk assessments, these economic data should be combined with probability assessments of future climatic extremes from global and regional climate models, as well as models of smaller-scale phenomena such as hurricanes[6] or tornadoes. Other natural-hazard models might be included.

Understanding the global economic network is a huge challenge. The harmonization and cross-checking of the information will require constant effort, and some inconsistencies between data on different scales will be unavoidable. At the same time, the more information that is added, the better the network will get.

Although we will never be able to predict every impact, we must aim to provide the best information we can because society needs to decide what to do, even in the presence of uncertainty[7]. With the wrong focus, we will protect the wrong places with the wrong tools. ■

Anders Levermann *is professor of dynamics of the climate system at the Potsdam Institute for Climate Impact Research, Germany; and is at the Institute of Physics, Potsdam University, Germany.*
e-mail: anders.levermann@pik-potsdam.de

1. Intergovernmental Panel on Climate Change *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (IPCC, 2013).
2. Hallegatte, S. *Modeling the Roles of Heterogeneity, Substitution, and Inventories in the Assessment of Natural Disaster Economic Costs* (World Bank, 2012); available at http://go.nature.com/mtbynm.
3. Weitzman, M. L. *Rev. Econ. Stat.* **91,** 1–19 (2009).
4. World Bank *Thai flood 2011: Rapid Assessment for Resilient Recovery and Reconstruction Planning* (World Bank, 2012); available at http://go.nature.com/2xf1na.
5. Lenzen, M. *et al. Environ. Sci. Tech.* **46,** 8374–8381 (2012).
6. Emanuel, K. *Proc. Natl Acad. Sci. USA* **110,** 12219–12224 (2013).
7. Schellnhuber, H. J. *et al. Turn Down the Heat: Climate Extremes, Regional Impacts, and the Case for Resilience* (World Bank, 2013); available at http://go.nature.com/8iwsbh.



TOM STOCKILL/CAMERA PRESS LONDON

Svante Pääbo holds a cast of a Neanderthal's skull.

EVOLUTION

The human puzzle

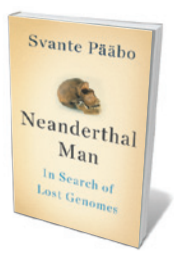
Henry Gee relishes the memoir of Svante Pääbo, a leader in the field of ancient DNA.

At a Royal Society meeting in London last year, just weeks before the publication in these pages of a high-quality Neanderthal genome (K. Prüfer *et al. Nature* **505,** 43–49; 2014), David Reich — one of the paper's authors — spoke of "introgression" between Neanderthals, *Homo sapiens* and other hominins. This irked a member of the audience. "Are you telling me," he asked, in cut-glass tones, "that these different species copulated with one another?" I was seized by an impulse to stand up and reply, in similarly stentorian fashion, "Not only did they copulate, but their union was blessed with issue!" (I stayed in my seat.)

The study of human origins and evolution currently stands on a cusp. For decades we have had to make do with bones and stones, thin gruel from which to craft a narrative. Now we can extract DNA from fossils. Not just in bits and pieces, each as enigmatic as a broken tooth or a chipped stone flake — but entire genomes. Unlike fossils, genomes can tell stories. They can legitimately link species into skeins of common ancestry and descent.

➲ NATURE.COM
For more on Neanderthals, see:
go.nature.com/do74np

If there is one name associated with ancient DNA, it is Svante Pääbo. Now at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, Pääbo pioneered and has largely led the field for the past three decades. His book, *Neanderthal Man*, is perfectly timed, beautifully written and required reading — it is a window onto the genesis of a whole new way of thinking. (I should add a disclaimer at this point. I have a walk-on part in *Neanderthal Man*. Pääbo is as disarmingly candid about journals and editors as he is about anything else. I get off lightly.)



Neanderthal Man: In Search of Lost Genomes
SVANTE PÄÄBO
Basic Books: 2014

The book is primarily a memoir. Pääbo recounts his life story with a Fennoscandian frankness that some readers might find disconcerting. Along the way, he tells us a great deal about science and scientists. There is mercifully little of the didactic treatment of the structure of DNA and genes that authors feel obliged to rehearse on such occasions. Dispensing quickly with such banal necessities, Pääbo gets on with the cutting-edge science to which he was witness, and in some cases helped to create — the astonishing development of devices that could be used

to sequence DNA ever more efficiently and at lower and lower cost. He describes the technology clearly, almost like a recipe book: you feel you should have *Neanderthal Man* on the bench as you try its techniques for yourself.

Thanks to these developments, scientists are finding many more species of extinct hominin lurking out there in the shadows, betrayed by no known fossil evidence. For example, Denisovans, extinct hominins that lived in Siberia until relatively recent times, are much better known from their DNA than from the tally of their fossils — a small, nondescript finger bone and a peculiar tooth. And yet in their DNA are traces of yet another unknown species — glimpsed only as stretches of nucleotides — as evanescent as the smile of the Cheshire Cat.

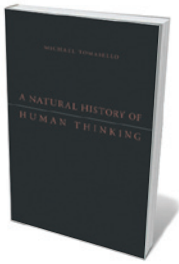
> ***"You feel you should have* Neanderthal Man *on the bench as you try its techniques for yourself."***

Pääbo illustrates how the advent of ancient DNA has already had a profound effect on our understanding of human evolution. Skulls and skeletons, once put away in cupboards lest they frighten the undergraduates, are being brought out into the light. Some of these peculiar specimens — such as the skull from Iwo Eleru in Nigeria that looks archaic but is only 13,000 years old — may represent evidence of a richer and much more diverse prehuman history than we are used to thinking about. It has taken the recovery of ancient DNA, not more fossil bones, to jolt us into this wider reality, to force our gaze over a great, unexplored new world.

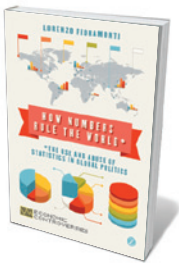
But as Pääbo recounts, there have been many false positives along the way. He deals harsh judgement on some of the grand claims from the Wild-West phase of ancient DNA research, before secure protocols had been established (and no, we at *Nature* don't escape his searchlight glare). And he does not spare himself from criticism. He looks back on the beginnings of his career in the 1980s, when, torn between a fascination for Egyptology and biochemistry, he mixed the two and tried to extract DNA from an Egyptian mummy. He thought he was making history. What he made was a mess. But, like all true scientists, he never gave up, finding all sorts of ways to achieve his goals, inventing new techniques and new ways of seeing. Eventually, in 1985, he reported the successful cloning of DNA from a mummy, and history was made. ■

Henry Gee *is a Senior Editor of* Nature *and the author of* The Accidental Species: Misunderstandings of Human Evolution.

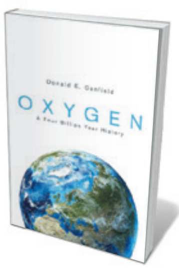
Books in brief



A Natural History of Human Thinking
Michael Tomasello Harvard University Press *(2014)*
In this prequel to his 1999 *Cultural Origins of Human Cognition* (Harvard University Press), developmental psychologist Michael Tomasello argues that human thinking is unique because it is cooperative. He posits that environmental upheavals forced early humans to channel their thinking towards collective aims through two evolutionary innovations: collaboration while foraging, and the rise of culture as population and competition burgeoned. Tomasello convincingly sets out how "shared intentionality", in which social complexity spawned conceptual complexities, sets us apart.

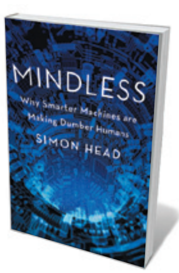


How Numbers Rule the World: The Use and Abuse of Statistics in Global Politics
Lorenzo Fioramonti Zed Books *(2014)*
Globally, we love statistics. Indexes and indicators produced by social-science bodies alone number in their hundreds, providing grist for policy mills around the world. In this intelligent study of pervasive quantification, Lorenzo Fioramonti questions its grip on society. Numerical reasoning in overdrive, he argues, can create distorted pictures of real life, amplify the power of markets and sap debate. Packed with cogent analyses of everything from credit-rating agencies to the manipulation of statistics by climate sceptics.



Oxygen: A Four Billion Year History
Donald E. Canfield Princeton University Press *(2014)*
Ecologist Donald Canfield delivers an engaging and authoritative primer on oxygen, that vital element comprising more than one-fifth of our atmosphere. In tracing its 4-billion-year history, Canfield proffers cutting-edge findings on geological and biological questions from deep time. He explores Earth's 'Goldilocks' status; squeezes into the *Alvin* deep-diving submersible to muse on life before oxygen; and probes photosynthesis, the rise of oxygenating cyanobacteria, stabilization of atmospheric oxygen, the 'great oxidation event' 2.4 billion years ago, and the ancient links between organisms and O_2.

Cybersecurity and Cyberwar: What Everyone Needs to Know
P. W. Singer and Allan Friedman Oxford University Press *(2014)*
The pace of global digitization, and the widespread lack of understanding of related security risks, is a ticking time bomb. Thus argue P. W. Singer and Allan Friedman in this broad-ranging overview of cybersecurity. They start with basics such as software vulnerabilities, then delve into the implications of and solutions to security breaches, touching on hot issues such as resilience and the controversial use of overlay systems that endow online anonymity, such as Tor. If you don't know your asymmetric cryptography from your spear phishing, this is a thoughtful introduction.



Mindless: Why Smarter Machines are Making Dumber Humans
Simon Head Basic Books *(2014)*
'Computer business systems' (CBS) are increasingly embraced in finance, business and health care to monitor the performance of employees digitally — to pernicious effect, argues Simon Head. At a time of deepening inequalities in wealth, he writes, such complex digital control of workplace behaviour disempowers those who can ill afford it. Head presents compelling examples of the impacts of CBS at Goldman Sachs, Amazon and Taiwanese electronics manufacturer Foxconn, among others. **Barbara Kiser**

GEOLOGY

Bedrock of China

Xu Xing applauds a study tracing the links between Chinese nationalism and geology.

Unearthing the Nation: Modern Geology and Nationalism in Republican China
GRACE YEN SHEN
University of Chicago Press: 2014.

Chinese science has long been tightly entangled with nationalism. An illuminating case study is the development of geology during the Republican era (1911–49). This followed an unusual pattern, striking a balance between the interests of science, the nationalist movement, the state and scientists in difficult, unstable circumstances. Science historian Grace Yen Shen chronicles the field's evolution in *Unearthing the Nation*.

Shen begins with an account of foreign exploration in Chinese territory from the mid-nineteenth to the early twentieth centuries, such as US geologist Raphael Pumpelly's investigations of the coalfields near the Yangtze River in the 1860s, and German geologist Ferdinand von Richthofen's field trips across China not long after. Richthofen went on to publish milestone works such as the five-volume *China: The Results of My Travels and the Studies Based Thereon* (1877–1912). In the early twentieth century, Chinese researchers, including the German-trained Gu Lang and Zhou Shuren, published on geology themselves.

Zhou (who under his pen name Lu Xun is a giant of Chinese fiction) was the first Chinese person to write on the field, in *Brief Outline of Chinese Geology* (1903). But as Shen notes, it was the investigations of Zhang Hongzhao, Ding Wenjiang, Weng Wenhao, Li Siguang and others around this time that marked the first stirrings of a homegrown discipline. Weng became the first Chinese geologist to earn a doctorate, after investigating the igneous formations of Belgium for his thesis at the University of Louvain. These pioneers, Shen says, saw fieldwork as helping China to "understand its own territory": science thus became a means of nation-building.

Yet for years, Chinese geology remained internationally collaborative in terms of practitioners, fieldwork, institutions and publications. In the 1920s, China was primarily agrarian and lacked the financial and intellectual resources to cultivate science. The Geological Society of China (GSC), established in 1922, was the first scientific association initiated by Chinese investigators. It listed among the 78 members in its first year 23 foreigners — including Swedish geologist Johan Gunnar Andersson, who contributed to the discovery of the Peking Man *Homo erectus* fossils. The *Bulletin of the Geological Society of China*, launched in 1922 and one of the first technical journals dedicated to Chinese geology, was published mainly in Western languages, including English. US geologist Amadeus Grabau (1870–1946), who spent most of his academic life in China, made huge contributions to Chinese palaeontology and stratigraphy, and the New York-based Rockefeller Foundation sponsored organizations such as the Cenozoic Research Laboratory in Beijing, established in 1928 to investigate the Peking Man fossils.

Chinese geology students on a field trip in about 1950.

MARC CHARMET/THE ART ARCHIVE

Chinese geologists persisted in fostering an independent discipline, even in 1927–37, when frequent conflicts flared between the government in Nanjing and local warlords, and within the ruling party. Weng and others recognized that their field could help to satisfy practical needs of the state such as the search for fossil fuels, and could build national pride. A platform came in 1936 with the GSC's Chinese-language journal *Dizhi Lunping* (*Geological Review*). And the Second Sino-Japanese War of 1937–45 was a watershed: the drive to find natural resources for the war effort led to achievements such as the discovery of China's first oil fields. Towards the end of the Republican era, a truly Chinese geological community had come together.

Shen's chronicle reveals a broader trend in Chinese science. In the 1930s, Weng and several other foundational Chinese geologists became high-level government officials. The desire of Chinese intellectuals to build a great nation has often led outstanding researchers into administration and politics, a tradition reflected in the saying '*Xue er you ze shi*' (Officialdom is the natural destination for good scholars'). The trend persists; in the long run I feel that it will harm Chinese science.

Unearthing the Nation is more than a scientific history. Shen's in-depth analysis reveals that national, political and cultural loyalties had a key role in the development of Chinese geology, and she seamlessly integrates this into her narrative on the discoveries and evolution of the field. Shen includes Chinese characters in the text, which makes the book more congenial for those who can read Chinese, and adds colour for those who cannot.

I would have loved to see more information on specific scientific discoveries, and Shen's tendency to focus on a limited number of key geologists and organizations sometimes obscures the larger picture. Nevertheless, this is an important book: it presents a comprehensive history of Chinese geology while demonstrating the discipline's unique pattern of development. Implicit in it is the significance of openness to international community, even in the development of a national scientific discipline. ■

Xu Xing *is a professor in the Institute of Vertebrate Paleontology and Paleoanthropology of the Chinese Academy of Sciences in Beijing.*
e-mail: xu.xing@ivpp.ac.cn

CHRIS FOSTER/MANCHESTER SCIENCE FESTIVAL 2013

Trevor Cox bursts a balloon under a railway arch in Manchester, UK, to demonstrate the resulting echo.

Q&A Trevor Cox

The sound hunter

Acoustical engineer Trevor Cox has designed concert halls, but recently turned to 'sound tourism' — gathering audible phenomena worldwide for his book Sonic Wonderland. *He talks about burping sand dunes, the bass baritone of a cracking glacier and the hiss of the nervous system.*

ILLUSTRATION: NICK HIGGINS; ORIGINAL PHOTOGRAPH: UNIV. SALFORD

How did you get into sound tourism?
A few years ago in London I went down a Victorian sewer and heard this amazing spiralling echo. It made me wonder what other curious sounds might be out there. I decided to turn my training in acoustic engineering on its head and seek out aural distortions and illusions. Travel guides to the sights rarely said much about sounds, and I couldn't find a list of the most amazing sounds in the world. So I started a blog mapping unusual sonic spaces, and that became *Sonic Wonderland*.

What happens with noise in a chamber?
Whenever you make noise in an enclosed space, millions of reflections bounce around the room. Echo is when you hear a separate reflection, such as in a bad concert hall where the trumpets sound like they are coming both from the stage and from the wall behind you. There is a mosque in Iran where, when you flick a piece of paper, the echo bounces between the ceiling and the floor seven or eight times. When your brain lumps room reflections into one event, that is reverberation. It adds a subtle bloom; without it your voice sounds dry and muffled.

What are the most reverberant spaces you have visited?
A mausoleum in Scotland claims to be the most reverberant place in the world, but it didn't seem that impressive on my visit. Then I learned about Inchindown in northwest Scotland, a depot built to protect fuel during the Second World War. The oil tanks are the size of enormous cathedrals dug into the side of a mountain, with thick concrete walls and no doors, and to get in you have to go through the pipework. You can have a quiet conversation in there because the walls are so far away. But as soon as you raise your voice this fog rises around you, a haze of echoes that build up and resonate for about a minute and a half.

How does noise travel over large distances?
In the early nineteenth century, sailors off the coast of Brazil reported hearing the sound of bells ringing some 160 kilometres away. That may have been due partly to the concave shape of the ship's sail, which might have reflected the sound, and to a wide layer of cold air over the ocean that might have refracted the sound back downwards to the ship. For similar reasons, on some nights when the weather is right, I can hear the crowd at Manchester United's football ground quite clearly from my house, even though it is more than three kilometres away. Inside buildings, it has been known for centuries that a curved ceiling can transmit sound across a large room. In some cathedrals, if you whisper into the walls, the sound will skim the dome and come through clearly tens of metres away.

What about the sounds of ice?
Ice makes so many sounds. You can get the most catastrophic bass notes when great chunks of a glacier drop off into the ocean. If you throw rocks on to a frozen lake, it sounds a bit like phaser gunfire from George Lucas's *Star Wars* films. When bits of ice wash up around the shore of a glacial lagoon, you get a gentle tinkling sound like wind chimes. There is a musician in Norway, Terje Isungset, who makes trumpets and xylophones out of ice. He has to source it from lakes that froze slowly, to ensure a regular crystalline structure. He calls them "the only instruments you can drink after you've finished playing".

And sand dunes?
Explorers Marco Polo and Charles Darwin observed that some sand dunes make rumbling sounds when you walk on them, owing to the uniform size of the grains and whether they are loose and sifted. Scientists describe the sound as you walk on the dune as burp-like, but to me it sounds more like a tuba. If you scoot down the dune on your rear, you can get a couple of metres of sand to vibrate. With more people, more of the dune surface vibrates and that creates a huge avalanche of sound, a continuous booming that can travel for a kilometre or so.

Is total silence possible?
Before I wrote the book, my answer would have been no. When I work in my anechoic chamber, a room that deadens most sound, I have found that you cannot get rid of bodily sounds such as blood pumping through your head or, if you are unlucky, a hissing that is probably spontaneous firings on the auditory nerve, a bit like tinnitus. But while researching the book I went to some places — such as a sensory-deprivation flotation tank and a remote peat bog in Northumberland, UK — where I was not conscious of any sound whatsoever. My suspicion is that, for some continuous sounds like the hissing of the nervous system, your brain just learns to ignore it after a while. ■

INTERVIEW BY JASCHA HOFFMAN

Correspondence

Medical data: the choice to opt out

You accuse the National Health Service (NHS) in England of using "sleight of hand" in the way we are advertising the care.data programme (see go.nature.com/srp5nu), suggesting that we should make it clearer to people that the programme poses potential risks to their privacy and that they can opt out of it (*Nature* **505,** 261; 2014). We believe that this accusation is unwarranted.

"You have a choice" is written in bold on the cover of the leaflet about the programme, which is being sent to every household in the country. The leaflet goes on to say: "If you do not want information that identifies you to be shared outside your GP [general practitioner] practice, please ask the practice to make a note of this in your medical record."

Last month, we published a detailed assessment of the potential negative and positive impacts of the programme on privacy (see go.nature.com/xcqaql). And, most importantly, patients have the opportunity to discuss the changes with a trained adviser and with their GP.

It would be unethical to introduce this opt-out system without proper publicity, as well as illegal under the UK Data Protection Act 1998. This accounts for the scale of our awareness-raising strategy and our advice last August to all GP practices to start telling people about the proposed changes.

Geraint Lewis *NHS England, Leeds, UK.*
geraint.lewis@nhs.net
Competing financial interests declared: see go.nature.com/sluxqa for details.

Medical data: widen use in research

The Wellcome Trust and other UK medical-research charities support the plans of the National Health Service (NHS) in England to make better use of information from patients' records, but we have no wish to downplay the right of people to opt out of the NHS care.data programme (go.nature.com/srp5nu), as you imply (*Nature* **505,** 261; 2014). Like you, we believe it is critical that the risks, benefits and choices are explained clearly to everyone.

We have launched a campaign to support the wider use of medical records for research through mechanisms such as the Clinical Practice Research Datalink, rather than the care.data programme specifically (see www.patientrecords.org.uk). It is intended to complement NHS England's communications by highlighting the choices people have alongside the research benefits we perceive, and to help people to reach an informed decision.

Those with concerns about sharing patient data are right in that no system can guarantee protection against determined misuse. We have confidence, however, in the strict safeguards that govern the research use of medical records, which can manage those risks while enabling research to benefit from a national cradle-to-grave data set.

Jeremy Farrar *Wellcome Trust, London, UK.*
j.farrar@wellcome.ac.uk
Competing financial interests declared: see go.nature.com/lskrj4 for details.

Planck team replies to data 'anomalies'

We would like to clarify some points arising from your News report on the debate over data from the European Space Agency's Planck mission (see *Nature* http://doi.org/q8t; 2013).

The cosmological parameters estimated by the Planck Collaboration are statistically compatible with those estimated by NASA's Wilkinson Microwave Anisotropy Probe (see G. Hinshaw *et al. Astrophys J. Suppl. S.* **208,** 19; 2013). Also, the analysis of the Planck data by David Spergel and colleagues (see preprint at http://arxiv.org/abs/1312.3313*;* 2013) is actually in close agreement with our own (http://arxiv.org/abs/1303.5076; 2013): the values of their parameters are within one standard deviation of ours.

For example, their value of the Hubble constant is within 0.6 of a standard deviation of ours; the matter density and the amplitude of the fluctuation spectrum differ by about one standard deviation. These differences, which are not evident in our analyses of the Planck data, could be caused by methodological variations between the respective analyses rather than by systematic errors in the Planck data.

We, and Spergel and colleagues, have verified that the small, time-dependent systematic errors that affect a subset of the data at a radio frequency of 217 gigahertz, which we reported on in the revised versions of the Planck papers from 2013, have little impact on the Planck Collaboration's cosmological results.

Jan Tauber *European Space Agency, Noordwijk, the Netherlands, and the Planck Science Team (see go.nature.com/q5ltry), on behalf of the Planck Collaboration.*
jtauber@rssd.esa.int

Carbon dioxide storage is secure

The Sleipner gas field in the North Sea has the world's first purpose-engineered subsea geological storage site for carbon dioxide. Contrary to your headline's implication, seabed fractures do not pose any threat to this project (*Nature* **504,** 339–340; 2013).

Independent researchers have analysed extensive data from site monitoring using seismic-reflection surveys of the deep subsurface (both before CO_2 injection and then at two-year intervals); they found that performance is excellent, with no evidence of any CO_2 leakage (see A. J. Cavanagh and R. S. Haszeldine, *Int. J. Greenh. Gas Con.* **21,** 101–112; 2014).

Your graphic, which juxtaposes stored CO_2 with fractures, is also misleading: Sleipner is in fact 25 kilometres away from the fracture described and is overlain by 500 metres of sealing mudrock from the estimated depth of the crack. Elsewhere beneath the North Sea, mudrocks have retained natural CO_2 for tens of millions of years.

The suggestion that leakage would be "a disaster for public opinion" is unsupported. Social-science research indicates that unintended leakage need not be a show-stopper (see L. Mabon *et al. Mar. Policy* **45,** 9–15; 2014). More than guarantees that sites will never leak, the public seeks reassurance that site selection minimizes leakage risk, and that monitoring and remediation procedures are in place should a leak be discovered.

There are many known fluid conduits beneath the North Sea, but there is no evidence of unplanned CO_2 or methane movement in the rocks overlying the storage site. Since the Sleipner project was set up 20 years ago, global endeavours have improved the geoscientific identification, operation and monitoring of CO_2 storage (see V. Scott *et al. Nature Clim. Change* **3,** 105–111; 2013). Sleipner's CO_2 is securely retained by residual saturation in the reservoir, multiple mudrock seals, and eventual dissolution and dispersion in pore waters.

Vivian Scott* *Edinburgh University, UK.*
vivian.scott@ed.ac.uk
**On behalf of 6 co-signatories (see go.nature.com/zivosz for full list).*

CONTRIBUTIONS

Correspondence may be sent to **correspondence@nature.com** after consulting the guidelines at **go.nature.com/cmchno**. Alternatively, readers may comment online: **www.nature.com/nature**.

John Cornforth

(1917–2013)

Nobel-prizewinning chemist who tracked how enzymes build cholesterol.

BETTMANN/CORBIS

Life depends on the geometric intricacies of enzymatic reactions. Even when molecules are exact mirror images of each other, enzymes treat the 'left-handed' and 'right-handed' versions differently. John Cornforth identified which of a series of mirror images interact with the enzymes that carry out the natural synthesis of cholesterol. This work, for which he shared the 1975 Nobel Prize in Chemistry, laid the foundations for many studies of how cells build organic compounds.

Cornforth, who died on 8 December 2013, was born in Sydney, Australia, in 1917. By the time he was ten, the first signs of his oncoming deafness had become apparent. As a boy, he built his own rudimentary laboratory at home. And, encouraged by a school teacher, he entered the University of Sydney at the age of 16 to read chemistry, a subject in which he thought his deafness would be less of a handicap. Although unable to hear the lectures, his thorough study of the scientific literature enabled him to graduate in 1937 with a first-class honours degree and a university prize.

Boyhood rambles in the bush inspired Cornforth's interest in natural products, and he began graduate studies at the University of Sydney. A number of his early papers were on the constituents of Australian plants, such as the caustic vine (*Sarcostemma australe*). His lifelong nickname, Kappa, arose from chemists' habit at the time of engraving their glassware: his initials (JC) resembled the Greek letter.

Cornforth's deafness led to an intense loneliness that was alleviated by the companionship in the laboratory. The skills he developed in his home lab, of building and repairing experimental apparatus, had many benefits. One was meeting the talented chemist, Rita Harradence, who asked him to repair a flask. In 1941, she became his wife. Throughout his career she acted both as an interpreter and a collaborator; they authored more than 40 papers together.

In 1939, Cornforth and Harradence were awarded scholarships for doctoral studies under Robert Robinson, an organic chemist at the University of Oxford, UK, who won the 1947 Nobel Prize in Chemistry. They began work on the synthesis of steroids, a biologically important class of complex, multi-ringed organic compounds that includes cortisone, estrone and testosterone. This effort eventually bore fruit in the first total synthesis of an androgenic hormone, reported in 1953 (H. M. E. Cardwell *et al. J. Chem. Soc.* 361–384; 1953).

In 1942, as part of the joint US–UK war effort, the couple joined the team working on the structure of the antibiotic penicillin. Cornforth made a number of important contributions, including identifying and synthesizing penicillamine, a key degradation product of penicillin. This work stimulated Cornforth's investigations into the class of compounds known as oxazolones, including a type of chemical rearrangement that now bears his name.

In 1946, the Cornforths moved to the Medical Research Council National Institute for Medical Research in London. Here, Cornforth continued his work on steroid and oxazolone chemistry and began a very fruitful collaboration with the medical biochemist George Popják, which continued when they became co-directors in 1962 of Shell Research's newly set up Milstead Laboratory of Chemical Enzymology in Sittingbourne, UK.

Like many chemists, Cornforth was intrigued by how natural products are formed. In the years after the Second World War, the radioisotope carbon-14 became available for basic research, providing a way to establish the biological building blocks of larger molecules. Other researchers had already begun to figure out the origin of certain carbon atoms in a side chain of the cholesterol molecule by using radioactively labelled acetate, a small organic compound containing only two carbon atoms. Cornforth took on the more demanding experimental work required to establish the origin of each of the carbon atoms in cholesterol's four conjoined molecular rings.

He identified 14 steps in the early stages of the natural formation of cholesterol. In each of these steps, the intermediate products could be transformed in one of two ways. His design for labelling experiments defined a single pathway out of the 16,384 (2^{14}) possibilities.

In another series of experiments, on acetic acid, Cornforth labelled the hydrogen atoms around a carbon, replacing the hydrogens with the isotopes deuterium and tritium such that each had a distinct position around carbon. These classic experiments opened up the possibility of exploring a wide range of enzyme reactions, including fatty-acid biosynthesis.

In 1975, the same year that he won the Nobel prize for decoding the stereochemistry of biosynthetic reactions, Cornforth accepted a Royal Society research professorship at the University of Sussex, UK. There, he began an extremely ambitious project to craft a compound that could act as an analogue for hydratase, the enzyme that adds water to another molecule.

He was knighted in 1977 and made a Companion of the Order of Australia in 1991. Kappa lectured undergraduates and supervised student projects until he was well into his 80s, often enhancing conversations with an aptly worded limerick. His kindness, generosity and humour were appreciated by all with whom he came into contact. ■

Jim Hanson *is professor emeritus of chemistry at the University of Sussex, Brighton, UK. He worked in the same laboratory as John Cornforth for three decades.*
e-mail: j.r.hanson@sussex.ac.uk

NEWS & VIEWS

FORUM: Microbiology

A talented genus

Members of a newly described candidate bacterial genus, *Entotheonella*, have been identified as the sources of the rich array of natural products found in the marine sponge *Theonella swinhoei*. Two scientists discuss this discovery from the perspectives of microbial ecology and drug discovery. SEE ARTICLE P.58

Hidden depths

MARCEL JASPARS

In this issue, Wilson *et al.*[1] describe the discovery of two new bacterial species with large genomes and rich biosynthetic repertoires. This combination is so rare that the new phylum to which they have been assigned might be heralded as the successor to the Actinobacteria, the phylum responsible for many of the world's antibiotics and anticancer agents. How did their discovery come about? By studying sponges: organisms identified by the pioneers of marine natural-product chemistry as the source of unparalleled chemical diversity.

This identification raised questions about how sponges produce such a range of compounds and what their roles might be. The variety of chemical reactions observed seemed too broad to be produced by a single organism, and sponges collected from different locations had different and often non-overlapping metabolite profiles. It was only when it was noticed that similar compounds could be found in organisms as divergent as sponges and beetles that a common, microbial origin was suggested[2].

Early work to confirm this idea involved blending and centrifuging sponges to separate cell populations, which revealed that the microorganisms living in the sponges had biosynthetic repertoires distinct from those of the sponge cells[2]. Subsequent studies tried to get a clearer picture of which organism produced which compound, but the possibility of compounds moving from the true producer to other organisms obfuscated a clear interpretation[3].

Despite this, evidence mounted that the producer was a bacterium named *Candidatus* Entotheonella palauensis[4] (*Candidatus* indicates that the bacterium had not yet been cultured). Subsequent comparisons showed that the pathways responsible for producing the related compounds pederin (isolated from the *Paederus* beetle) and theopederin A (isolated from the sponge *Theonella swinhoei*) were highly similar and probably came from a bacterium associated with the sponge and the beetle[5,6]. However, this combined evidence, although suggestive, did not quite complete the loop between microorganism, biosynthetic genes and chemistry.

Wilson and co-workers' study finally closes these gaps in our understanding and, in doing so, reveals hidden depths of biosynthetic capacity in a candidate phylum that they name Tectomicrobia (from the Latin *tegere*, to hide, to protect). The authors combined previous experimental separation methods with whole-genome sequencing of candidate organisms to assess the number and range of biosynthetic gene clusters present in members of the phylum's only genus discovered so far, *Entotheonella*. There is now incontrovertible evidence that *T. swinhoei* is host to this genus, and that *Entotheonella* species have large genomes (greater than 9 megabases), of which a high proportion is dedicated to natural-product biosynthesis (Fig. 1).

The authors assigned gene clusters to the biosynthesis of several compounds identified in *T. swinhoei* extracts, including onnamides/theopederins, polytheonamides, keramamides/orbiculamides and cyclotheonamides, and identified a further 24 biosynthetic clusters with predicted or unknown products. There is little apparent overlap in biosynthetic repertoire between the two *Entotheonella* species the authors have so far described, indicating a vast potential for new chemistry in this phylum. Thus, it seems that members of Tectomicrobia are talented producers of chemical diversity, similar to the Actinobacteria and Cyanobacteria, which both include species with large genomes and many biosynthetic gene clusters. It also seems that Tectomicrobia are widespread: Wilson *et al.* analysed 37 taxonomically diverse sponge species from 20 locations, including some from geographically distant regions, and found *Entotheonella* species in 28 of the samples.

There is little apparent overlap in biosynthetic repertoire between the two species, indicating a vast potential for new chemistry in this phylum.

This study shows that new biosynthetically talented microorganisms can be discovered, and suggests that systematic searches will yield further species in this phylum, as well as new phyla. Questions that remain include whether marine sponges are the only hosts for this phylum or whether it is more widespread; what the benefit is to the sponge of hosting such a talented symbiont; and how its presence in the sponge is controlled.

Marcel Jaspars *is at the Marine Biodiscovery Centre, Department of Chemistry, University of Aberdeen, Old Aberdeen AB24 3UE, UK.*
e-mail: m.jaspars@abdn.ac.uk

Supply and source

GREG CHALLIS

Bioactive natural products isolated from sponges and other marine animals offer interesting possibilities for treating cancer and other diseases. However, obtaining sufficient quantities of such metabolites from the marine environment for clinical trials is challenging. Wilson and colleagues' identification of bacteria from the candidate genus *Entotheonella* as the producers of most of the metabolites isolated from *T. swinhoei* suggests new approaches for overcoming this supply problem.

Natural products have diverse applications in medicine and agriculture. Iconic examples include penicillins and cephalosporins, used to treat bacterial infections; the cancer drug paclitaxel (Taxol); artemisinin, which targets the malaria parasite; the cholesterol-lowering drug lovastatin; and the insecticide spinosyn. The overwhelming majority of such compounds are produced by plants or terrestrial microorganisms.

Although marine sponges are another important source of bioactive natural products, only a handful of sponge natural products have entered the market. This is due primarily to the supply problem. For example, considerable quantities of a drug candidate are required for clinical trials, but only a few milligrams of

A, TOSHIYUKI WAKIMOTO; B, TETSUSHI MORI

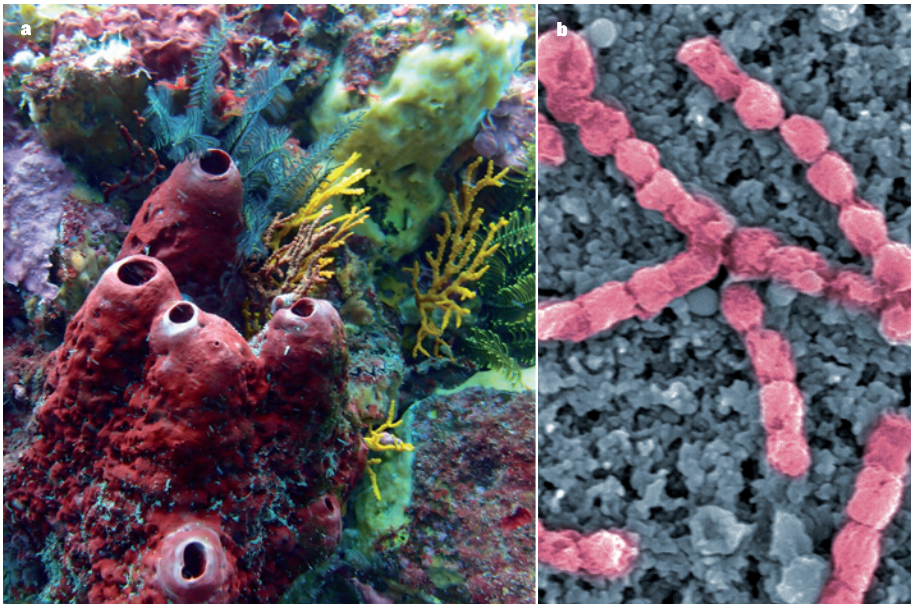


Figure 1 | Sponge's secret. Wilson *et al.*[1] show that many of the chemically diverse natural products found in the marine sponge *Theonella swinhoei* (**a**) are produced by biosynthetically talented bacterial symbionts (**b**; false-coloured), which they assign to the candidate genus *Entotheonella*.

most natural products can be isolated from the marine environment and it has hitherto proved impossible to cultivate the sponges from which such compounds are derived.

One approach to solving this problem is total chemical synthesis, which has been successfully used to produce the anticancer compounds discodermolide[7] and eribulin[8]. However, the structural complexity of most marine natural products means that developing efficient routes for their total synthesis is challenging. Another approach involves semi-synthesis from a structurally related metabolite. This has been applied[9] to the anticancer agent trabectedin, isolated from a sea squirt, which can be synthesized from safracin B produced by a cultivable terrestrial bacterium. But such routes are viable only if an abundant supply of an appropriate precursor is available.

Evidence has been mounting that uncultivated bacterial symbionts of marine sponges, rather than the sponges themselves, are the true producers of many bioactive metabolites[6,10]. However, it has been unclear whether several microbial inhabitants are responsible, or just one. Wilson *et al.* have answered this question, although it remains to be seen whether their report of *Entotheonella* species being responsible for producing the diverse array of metabolites isolated from the sponge is a widespread phenomenon among other sponges.

The authors' findings illuminate two promising approaches for addressing the supply problem. The first is large-scale cultivation of the microorganisms that produce interesting metabolites. This is likely to prove difficult, but the ability to obtain a draft genome sequence from a single microbial cell, as exemplified by Wilson and colleagues, may help to determine optimal culture conditions for the organisms. This comes with the caveat, however, that growing such microorganisms in pure culture might downregulate their production of bioactive metabolites — the biosynthetic pathways for similar metabolites in easily cultivable terrestrial microorganisms are often expressed poorly in pure cultures, or not at all, presumably because the environmental cues responsible for eliciting them are absent. Thus, genetic manipulation may be required to maintain desirable levels of metabolite production[11].

The second potential way to address the supply problem involves expressing the biosynthetic pathway of interest in an easily cultivable surrogate host. This synthetic-biology tactic has been used to produce a key intermediate of artemisinin biosynthesis in yeast[12]. Genome-sequence data may help to guide selection of the most appropriate surrogate, but extensive genetic manipulation will probably be required to optimize the production of each metabolite.

Wilson *et al.* also show that, as is the case for terrestrial bacteria such as *Streptomyces* species[13,14], *Entotheonella* species contain several pathways that hint at their ability to assemble previously unknown metabolites. This suggests that members of the genus might serve as a useful source of leads for drug discovery. ■

Greg Challis *is in the Department of Chemistry, University of Warwick, Coventry CV4 7AL, UK.*
e-mail: g.l.challis@warwick.ac.uk

1. Wilson, M. C. *et al. Nature* **506,** 58–62 (2014).
2. Bewley, C. A. & Faulkner D. J. *Angew. Chem. Int. Edn* **37,** 2162–2178 (1998).
3. Unson, M. D. & Faulkner D. J. *Experientia* **49,** 349–353 (1993).
4. Schmidt, E. W., Obraztsova, A. Y., Davidson, S. K., Faulkner, D. J. & Haygood, M. G. *Mar. Biol.* **136,** 969–977 (2000).
5. Piel, J. *Proc. Natl Acad. Sci. USA* **99,** 14002–14007 (2002).
6. Piel, J. *et al. Proc. Natl Acad. Sci. USA* **101,** 16222–16227 (2004).
7. Mickel, S. J. *et al. Org. Process Res. Dev.* **8,** 122–130 (2004).
8. Yu, M. J., Kishi, Y. & Littlefield, B. A. in *Anticancer Agents From Natural Products* (eds Cragg, G. M., Kingston, D. G. I. & Newman, D. J.) 317–346 (Taylor & Francis, 2005).
9. Cuevas, C. & Francesch, A. *Nat. Prod. Rep.* **26,** 322–337 (2009).
10. Freeman, M. F. *et al. Science* **338,** 387–390 (2012).
11. Laureti, L. *et al. Proc. Natl Acad. Sci. USA* **108,** 6258–6263 (2011).
12. Ro, D.-K. *et al. Nature* **440,** 940–943 (2006).
13. Bentley, S. D. *et al. Nature* **417,** 141–147 (2002).
14. Ikeda, H. *et al. Nature Biotechnol.* **21,** 526–531 (2003).

This article was published online on 29 January 2014.

CANCER

Interference identifies immune modulators

A broad *in vivo* screen of the effects of specific gene inhibition on the antitumour activity of immune cells in mice bearing melanomas has revealed potential targets for cancer therapy. **See Article p.52**

LARS ZENDER

Therapies designed to boost the immune system's response to tumours hold great promise for overcoming drug resistance in cancer. Advanced solid tumours inevitably develop resistance against currently available cytotoxic or molecularly targeted therapies, but durable responses have been observed following some immunotherapeutic treatments, leading to speculation that a cure for some patients could be possible. On page 52 of this issue, Zhou *et al.*[1] use RNA-interference technology to identify genes that can be targeted to enhance the robustness and proliferation of immune cells called $CD8^+$ T cells in mice bearing melanomas.

Two existing targets for cancer immunotherapy are the receptor molecules CTLA-4

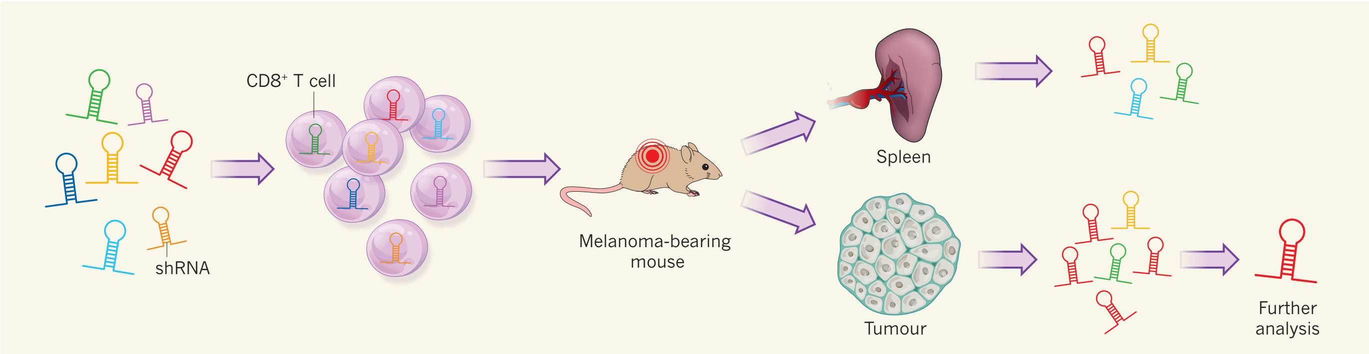


Figure 1 | Screening for T-cell modulators. Zhou *et al.*[1] compiled libraries of short hairpin RNA (shRNA) molecules designed to specifically inhibit the expression of genes expressed by dysfunctional (anergic or exhausted) $CD8^+$ T cells or of genes encoding cell-signalling enzymes (phosphatases and kinases). The authors then infected $CD8^+$ T cells with these shRNAs and implanted the cells into mice bearing melanomas. Seven days later, they isolated T cells from the spleens and tumours of the mice and compared the relative representation of shRNA molecules in the two tissues. Those shRNA molecules found to be enriched in the tumour were postulated to be involved in facilitating T-cell survival or proliferation within the tumours and were selected for further analysis.

and PD-1, which are expressed on the surface of T cells and transmit signals that dampen the cells' immune activity. Antibodies that bind these receptors, and thereby relieve the immune inhibition, have emerged as a powerful treatment for patients with advanced melanoma[2,3], and the anti-CTLA-4 antibody ipilimumab was the first immunotherapy shown to significantly improve the overall survival of these patients[4]. However, patient responses to tumour immunotherapy are highly variable, and it is unclear why some tumours respond and others do not. Thus, to improve the efficiency of such treatments further, a deeper understanding and better mechanistic characterization of antitumour immune responses are needed.

The discovery of RNA interference (RNAi) — the process by which small RNA molecules specifically inhibit gene expression by binding to and inducing the cleavage of messenger RNAs — has revolutionized loss-of-function genetic studies. The experimental application of RNAi, using collections of short-interfering RNA (siRNA) or short hairpin RNA (shRNA) molecules, means that screens of gene function can be conducted in almost every biological system. For example, shRNA screens have been used successfully to dissect tumour-suppressor gene networks, to identify modulators of drug resistance and to pinpoint vulnerabilities in cancer cells[5–8]. Furthermore, *in vitro* shRNA screening was recently applied to identify genes that regulate the differentiation of T cells into T-helper 1 and 2 subsets[9]. However, RNAi-based functional genetic screens are commonly performed *in vitro*, and as such do not take into account the effects of tumour microenvironment on cancer growth or immune-cell function.

Zhou and colleagues have taken shRNA screening in immune cells to the next level by building on advances in stable shRNA technology and *in vivo* shRNA screening[5–8,10]. The authors compiled two thematically focused libraries of shRNA molecules (Fig. 1). The first comprised 1,275 shRNAs targeting 255 genes whose expression is associated with T-cell exhaustion or anergy — states of functional inactivity that arise in cancer. The second contained 6,535 shRNAs targeting 1,307 genes encoding kinase and phosphatase enzymes involved in cell-signalling pathways. These shRNAs were then delivered by lentiviruses into activated mouse $CD8^+$ T cells carrying a specific T-cell receptor (OT-1); those cells that stably expressed shRNA molecules were then implanted into mice harbouring aggressive melanomas that expressed a model-antigen protein (Ova), which can activate the OT-1 T-cell receptor.

Seven days later, the OT-1 T cells were purified from the tumours and spleens of the mice and analysed to identify shRNA molecules that were substantially more highly represented in tumoural than in splenic T cells — the implication being that the genes targeted by these shRNAs are involved in mediating survival or proliferation of T cells in tumours. Top-scoring genes were then screened again using 15 different shRNAs targeting each candidate gene. In addition to several shRNAs targeting genes with known functions in T cells, the authors identified shRNAs against Ppp2r2d, a regulatory subunit of the phosphatase enzymes of the PP2A family, as strongly enriched in tumours.

In further experiments, the researchers found that shRNA-induced inhibition of Ppp2r2d expression increased the survival of $CD8^+$ T cells within the tumour and resulted in increased intratumoural $CD8^+$ T-cell proliferation. Most importantly, systemic delivery of melanoma-targeted $CD8^+$ T cells expressing an anti-Ppp2r2d shRNA resulted in increased death of melanoma cells, a significantly reduced tumour burden over time and prolonged survival of tumour-bearing mice. Strikingly, Ppp2r2d suppression not only improved the antitumour activity of $CD8^+$ T cells, but also increased that of another class of immune cell, $CD4^+$ T cells, thus suggesting a broad applicability of Ppp2r2d as a therapeutic target in T cells.

Zhou and colleagues' paper breaks ground on several levels. Their study sets a new standard in how the function of immune cells can be genetically dissected by RNAi screening *in vivo*. On the basis of the authors' data, similar screens in genetically engineered mouse models of different tumour types seem feasible and should be given high priority. Furthermore, functional screens specifically aimed at identifying modulators of $CD4^+$ T-cell function should be pursued.

The report also suggests an exciting potential for targeting Ppp2r2d in T cells for cancer therapy, assuming that the increased antitumour T-cell function observed following Ppp2r2d inhibition can be validated in other models — ideally ones that do not depend on model antigens and cell transplantation. It will also be interesting to explore the use of Ppp2r2d inhibition in conjunction with other immunotherapies. For example, Zhou *et al.* showed that shRNA-mediated suppression of Ppp2r2d did not reduce the expression of the inhibitory receptors PD-1 or LAG-3 on tumour-infiltrating T cells, so the effect of combining Ppp2r2d inhibitors with PD-1 or LAG-3 blockers should be investigated. ■

Lars Zender *is at the University of Tübingen, 72076 Tübingen, Germany, and at the German Center for Translational Cancer Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.*
e-mail: lars.zender@med.uni-tuebingen.de

1. Zhou, P. *et al. Nature* **506,** 52–57 (2014).
2. Hodi, F. S. *et al. N. Engl. J. Med.* **363,** 711–723 (2010).
3. Topalian, S. L. *et al. N. Engl. J. Med.* **366,** 2443–2454 (2012).
4. Robert, C. *et al. N. Engl. J. Med.* **364,** 2517–2526 (2011).
5. Zender, L. *et al. Cell* **135,** 852–864 (2008).
6. Berns, K. *et al. Cancer Cell* **12,** 395–402 (2007).
7. Meacham, C. E., Ho, E. E., Dubrovsky, E., Gertler, F. B. & Hemann, M. T. *Nature Genet.* **41,** 1133–1137 (2009).
8. Bric, A. *et al. Cancer Cell* **16,** 324–335 (2009).
9. Guo, L. *et al. Proc. Natl Acad. Sci. USA* **110,** E1849–E1856 (2013).
10. Wuestefeld, T. *et al. Cell* **153,** 389–401 (2013).

This article was published online on 29 January 2014.

ATMOSPHERIC SCIENCE

Drought and fire change sink to source

Aircraft have captured the 'breath' of the Amazon forest — carbon emissions over the Amazon basin. The findings raise concerns about the effects of future drought and call for a reassessment of how fire is used in the region. SEE LETTER P.76

JENNIFER K. BALCH

The Amazon forest accounts for 40% of the aboveground biomass stored in the world's tropical forests[1], but we do not know whether this crucial but threatened biome will be a sink or a source of atmospheric carbon in the coming decades[2]. Given the need to predict future climate scenarios, it is essential to refine our understanding of tropical forests' ability to sequester or release carbon[3]. The profiling of air columns over such forests by aircraft offers a much-needed window onto the major fluxes of tropical carbon. On page 76 of this issue, Gatti *et al.*[4] report the first estimate of carbon fluxes from the Amazon basin obtained in this way over the course of two years. Their findings suggest that the combined effects of drought and fires can cause the Amazon forest to become a net source of atmospheric carbon.

The authors sampled air masses several kilometres above the forest canopy at four Amazon locations, creating a patchwork of atmospheric profiles of carbon dioxide and carbon monoxide that spans the entire Amazon basin. They conducted these measurements during a major drought year (2010) and a relatively wet year (2011) for the region.

The researchers found that, during the drought year, burning of vegetation associated with land use and reduced photosynthesis resulted in 0.48 ± 0.18 petagrams of carbon (Pg C; 1 Pg is 10^{15} grams) being lost from the Amazon forest biome (uptake by the biome was 0.03 ± 0.22 Pg C per year; fire emissions were 0.51 ± 0.12 Pg C per year). During the wetter 2011, however, the Amazon was effectively carbon neutral: biome uptake (0.25 ± 0.14 Pg C per year) very nearly cancelled the fire emissions (0.30 ± 0.10 Pg C per year). Temperatures were above average in both years, but similar, suggesting that a moisture deficit reduced photosynthesis rates in 2010, rather than the crossing of a temperature threshold.

The growth rate of atmospheric CO_2 levels observed over the past five decades at Mauna Loa, Hawaii, and at the South Pole was recently shown[5] to be highly sensitive to year-to-year variability in tropical temperatures, and is further moderated by moisture conditions. This finding, taken together with Gatti and colleagues' study, implies that a shift in the terrestrial carbon cycle may be occurring because of the sensitivity to drought of tropical forests globally.

The world's vegetation takes up about 2.6 ± 0.7 Pg C per year, compared with around 9 Pg C per year emitted to the atmosphere, mostly as CO_2, from fossil-fuel combustion and cement production[6]. The Amazon forest accumulated an average of 0.4 Pg C per year in the two decades before 2005 — a range of 0.3–0.6 Pg C per year, estimated through repeated sampling of nearly 100 permanent plots across the basin[7] — and so has had a substantial role in offsetting global anthropogenic emissions of greenhouse gases. Whether this annual uptake will persist and compensate for emissions related to drought and land use in the future remains uncertain.

Gatti and colleagues' approach captures bi-weekly atmosphere–biosphere gas exchange across millions of square kilometres, the first time that this has been done at such a scale and for so long. Their method surpasses the spatial and temporal restrictions of, as well as some of the assumptions associated with, other methods such as plot-level inventories or modelling based on satellite data.

Atmospheric profiling using aircraft is a crucial tool in our understanding of Amazon carbon fluxes, and has the potential — if a pan-tropical network of aircraft observations can be established — to determine how tropical forests worldwide are responding to the combined threats of increasing drought and land-use pressures. A big advantage of this method is that it integrates emissions and uptake from naturally occurring land and river processes with land-use emissions to give a regional picture of total carbon fluxes. However, understanding the drivers and mechanisms behind these fluxes is key for the future management of carbon in tropical regions.

Given the importance of fire in shifting the Amazon basin from a sink to a source of carbon, one of the next steps is to reconcile the different fire types that contribute to the authors' regional estimates of fire emissions. Gatti and co-workers' vertical profiling detected carbon monoxide, which could have been caused by fires used for deforestation (Fig. 1), land management (pasture burning and 'slash-and-burn' agriculture, for example) and escaped understory wildfires[8]. More than 85,000 square kilometres of otherwise intact forests burned in understory fires in the southern Amazon during the 2000s, and, in dry years, the area affected can exceed the area deforested for agriculture and pasture[9]. These fires kill 8–64% of mature trees across Amazon forest sites[10], and burn biomass[11], thereby reducing forest carbon stocks. Teasing out the different land-use drivers that contribute to overall fire emissions is essential to aid fire-prevention and fire-management strategies that could help to reduce those emissions.

Because drought frequency and intensity in the Amazon might increase in the future[12], the authors' results are concerning. Furthermore, during the period of the study, deforestation rates were the lowest they had been since the records of Brazil's National Institute for Space Research began in 1988. The substantial fire emissions documented by the authors during their study therefore imply that efforts to reduce deforestation must also address the use of fire as a land-management tool. In sum, if drought and fire frequencies increase in the future, they may override the Amazon's function as a carbon sink. ■

JENNIFER K. BALCH

Figure 1 | Deforestation fire in the southeastern Amazon. Gatti *et al.*[4] report that a combination of severe drought and fires associated with land use can shift the Amazon region from being a sink to a source of atmospheric carbon.

Jennifer K. Balch *is in the Department of Geography, Pennsylvania State University, University Park, Pennsylvania 16802, USA.*
e-mail: jkbalch@psu.edu

1. Baccini, A. *et al. Nature Clim. Change* **2,** 182–185 (2012).
2. Davidson, E. A. *et al. Nature* **481,** 321–328 (2012).
3. IPCC *Climate Change 2013: The Physical Science Basis* (Cambridge Univ. Press, 2013).
4. Gatti, L. V. *et al. Nature* **506,** 76–80 (2014).
5. Wang, X. *et al. Nature* http://dx.doi.org/10.1038/nature12915 (2014).
6. Le Quéré, C. *et al. Nature Geosci.* **2,** 831–836 (2009).
7. Phillips, O. L. *et al. Science* **323,** 1344–1347 (2009).
8. Balch, J. K., Nepstad, D. C., Brando, P. M. & Alencar, A. *Science* **330,** 1627 (2010).
9. Morton, D. C., Le Page, Y., DeFries, R., Collatz, G. J. & Hurtt, G. C. *Phil. Trans. R. Soc. Lond. B* **368,** 20120163 (2013).
10. Barlow, J. & Peres, C. A. in *Emerging Threats to Tropical Forests* (eds Laurance, W. F. & Peres, C. A.) 225–240 (Univ. Chicago Press, 2006).
11. Balch, J. K. *et al. Global Change Biol.* **14,** 2276–2287 (2008).
12. Malhi, Y. *et al. Science* **319,** 169–172 (2008).

IMMUNOLOGY

Oiling the wheels of autoimmunity

Oily substances in the skin have now been shown to contain structures that activate a population of skin-homing, self-reactive T cells. The responses of these immune cells may contribute to local defences, but also to autoimmune disease.

MITCHELL KRONENBERG & WENDY L. HAVRAN

Immunology students are taught that the immune system responds to foreign entities while remaining tolerant to 'self' structures. This is not strictly true, however, because there are specialized populations of immune cells that are self-reactive. Such cells have the potential to initiate undesirable autoimmune reactions, so their existence raises several questions. What are the origin and structure of the self-antigens to which these cells respond, and how is this potentially dangerous self-recognition regulated? Reporting in *Nature Immunology*, de Jong *et al.*[1] identify hydrophobic self-antigens in the skin that are recognized in an unusual manner by a specialized subset of skin-resident immune cells.

B and T cells are the white blood cells responsible for immune recognition in the adaptive immune system. Populations of self-reactive T cells reside in or near the epithelial surfaces of the skin and intestine[2,3], where there are rich concentrations of microorganisms. Although it may seem paradoxical that self-reactivity is prevalent where microbes are abundant, it is possible that self-reactivity at these surfaces involves 'sentinel' immune cells that can rapidly respond to general signs of cellular stress or barrier disruption, without the need for specific recognition of microbes.

The antigens recognized by most T cells are peptides that are displayed on the surface of other cells, bound in the groove of antigen-presenting proteins of the major histocompatibility complex (MHC) family. Lipid antigens, by contrast, are bound by the hydrophobic grooves of CD1 antigen-presenting proteins, which are related to the MHC proteins. Humans have four CD1 proteins[4]: CD1a, CD1b, CD1c and CD1d. Some self and microbial lipid antigens that bind to CD1 proteins have been identified, but research on this antigen-presentation system has mostly been restricted to CD1d molecules.

De Jong *et al.* concentrated on T cells that recognize antigens bound to CD1a, which are more prevalent in human blood than T cells recognizing other CD1 proteins[5,6]. CD1a-reactive T cells are also found in the skin; when stimulated, these cells produce IL-22 (ref. 5), a cytokine protein involved in microbial defence and in inducing the proliferation of skin cells called keratinocytes. Moreover, Langerhans cells, which are antigen-presenting cells that reside in the skin's epidermal layer, express particularly high amounts of CD1a.

The authors show that a CD1a molecule purified from a human cell line activates CD1a self-reactive T cells by binding to their antigen receptor. The antigen-binding grooves of MHC and CD1 proteins are always filled, but the proteins do not discriminate self from non-self — this is the job of the antigen receptors that react to them. Therefore, the authors sought to uncover the CD1a-bound antigens that triggered the self-reactive T cells. Using mass-spectrometry analysis, they found more than 100 molecules corresponding in mass to

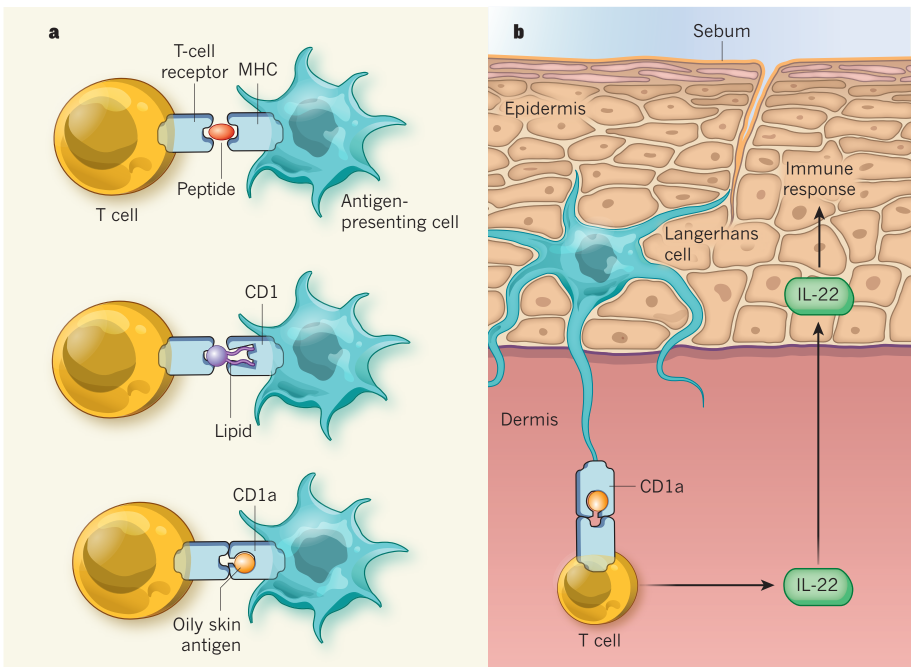


Figure 1 | Skin-antigen recognition by self-reactive T cells. a, T cells recognize peptide antigens bound to MHC molecules on the surface of antigen-presenting cells, or lipid antigens that are presented by CD1 proteins. In both cases, the antigen protrudes from the groove of the antigen-presenting molecule to engage the T-cell receptor. De Jong *et al.*[1] show that some T cells that react to CD1a molecules (a subset of the CD1 family) recognize oily substances found in sebum — a hydrophobic layer secreted onto the outermost layer of the skin. These self-antigens nestle deep within the CD1a groove, such that there might be direct contact only between CD1a and the T-cell receptor. **b,** The authors propose that skin-barrier disruption, through trauma or infection, allows Langerhans cells, which express CD1a, to acquire oily antigens from sebum and move from the skin's epidermis to the dermis. There, they make contact with self-reactive T cells and activate them to produce cell-signalling molecules, such as IL-22, that promote an immune response to barrier disruption without requiring specific recognition of invading microbes.

lipids (glycolipids and phospholipids) with different chain lengths and degrees of unsaturation. They then used several strategies to detect antigenic activity in this melange, including tests of synthetic molecules that had the same molecular masses as the material eluted from CD1a, and of partially purified lipids from various cell types.

The authors found that epidermal lipids bound to CD1a were more stimulatory for the CD1a-reactive T cells than lipids from other cell types. More specifically, they showed that CD1a-reactive T cells recognized highly hydrophobic compounds such as squalene and triacylglyceride, which are found naturally in the skin, when these were bound to CD1a. This reactivity was selective, and other hydrophobic molecules such as cholesterol were not recognized by the cells.

Although CD1 molecules have a hydrophobic antigen-binding groove, the completely hydrophobic character of the antigens presented by CD1a is surprising. T-cell antigen receptors typically recognize a composite structure of the antigen-presenting molecule and the antigen, with the exposed portion of the antigen participating in engaging the T-cell receptor (Fig. 1a). The exposed portion of lipid antigens is normally hydrophilic, containing a sugar or phosphate group, with the hydrophobic chains buried in the CD1 groove[7]. But the CD1a-binding self-reactive T cells did not obey this rule, because they did not require an exposed hydrophilic portion of the stimulatory lipid-containing antigen.

In fact, binding of lipids bearing hydrophilic groups to CD1a inhibited the response of these T cells, presumably by competing with more strongly hydrophobic antigens for binding into the CD1a groove. Therefore, it seems that the self-reactive T-cell antigen receptor requires a view of CD1a that is unimpeded by exposed hydrophilic groups; the bound lipid may simply permit or stabilize CD1a into the correct conformation. As a consequence, rather than recognizing single compounds with high specificity, these T cells can be stimulated by a range of highly hydrophobic substances that fit in the CD1a groove.

The skin forms a barrier to microbes through the generation of sebum — a highly hydrophobic substance synthesized in the sebaceous glands and secreted onto the outermost layer of the skin. Using microdissected sebaceous glands, de Jong *et al.* demonstrated that sebum is highly stimulatory for CD1a-dependent self-reactive T cells, and that it is rich in antigenic compounds, such as squalene. However, sebum is not typically in contact with the underlying dermal and epidermal layers that contain T cells and Langerhans cells. In normal skin, this physical separation may prevent CD1a-binding self-reactive T cells from being constantly exposed to their antigens. But disruption of the skin barrier by injury, infection or inflammation might allow sebum contents to permeate the epidermis and bind to CD1a-expressing Langerhans cells, thereby stimulating T-cell responses (Fig. 1b). Although this may aid general immune defences, in cases of prolonged barrier disruption, constant exposure of immune cells to sebum could contribute to autoimmune skin diseases such as psoriasis and atopic dermatitis.

Interestingly, squalene is currently used as an immune booster (adjuvant) to enhance the efficacy of vaccines and immunotherapies, and as a carrier for topical delivery to hair follicles of drugs for treating hair loss[8]. An autoimmune syndrome has also been described that is induced by adjuvants, including squalene[9]. It is possible that activation of CD1a-binding self-reactive T cells contributes to this compound's immune-stimulating effects. Certainly, further investigation of the regulation of these T-cell responses to skin oils is warranted, both for understanding immunity and autoimmunity and in light of the increasing therapeutic use of such agents. ■

Mitchell Kronenberg *is at the La Jolla Institute for Allergy & Immunology, La Jolla, California 92037, USA.* **Wendy L. Havran** *is at the Scripps Research Institute, La Jolla, California 92037, USA.*
e-mails: mitch@liai.org; havran@scripps.edu

1. de Jong, A. *et al. Nature Immunol.* **15,** 177–185 (2014).
2. Cheroutre, H., Lambolez, F. & Mucida, D. *Nature Rev. Immunol.* **11,** 445–456 (2011).
3. Witherden, D. A. & Havran, W. L. *J. Leuk. Biol.* **94,** 69–76 (2013).
4. Salio, M., Silk, J. D. & Cerundolo, V. *Curr. Opin. Immunol.* **22,** 81–88 (2010).
5. de Jong, A. *et al. Nature Immunol.* **11,** 1102–1109 (2010).
6. de Lalla, C. *et al. Eur. J. Immunol.* **41,** 602–610 (2011).
7. Zajonc, D. M. & Kronenberg, M. *Curr. Opin. Struct. Biol.* **17,** 521–529 (2007).
8. Aljuffali, I. A., Sung, C. T., Shen, F.-M., Huang, C.-T. & Fang, J.-Y. *AAPS J.* **16,** 140–150 (2014).
9. Vera-Lastra, O., Medina, G., Cruz-Dominguez, M. Del P., Jara, L. J. & Shoenfeld, Y. *Expert Rev. Clin. Immunol.* **9,** 361–373 (2013).

PARTICLE PHYSICS

Quarks are not ambidextrous

By separately scattering right- and left-handed electrons off quarks in a deuterium target, researchers have improved, by about a factor of five, on a classic result of mirror-symmetry breaking from 35 years ago. SEE LETTER P.67

WILLIAM J. MARCIANO

Symmetry makes the world go round. Scientific theories of the physics of elementary particles stem from simple symmetries that dictate the fundamental forces governing our Universe. Sometimes symmetries are broken, and that can have profound implications. An important case is the reflection, or right–left mirror, symmetry known as parity. On page 67 of this issue, an international team at the Thomas Jefferson National Accelerator Facility in Newport News, Virginia, reports[1] measurements of parity-symmetry breaking that confirm expectations and that unambiguously separate the electron and (much smaller) quark parity-violating interactions. The small quark parity violation can be used as a sensitive probe of new interactions or to measure subtle nuclear effects.

Elementary particles such as electrons and quarks (which make up protons and neutrons) carry intrinsic angular momentum called spin and act much like spinning tops. By convention, particles spinning clockwise with respect to their direction of motion are said to be left-handed, whereas their mirror images — those spinning anticlockwise — are right-handed. Parity symmetry swaps left and right, just as a mirror does.

Gravity, electromagnetism and strong nuclear forces all respect parity; that is, they are symmetrical (unchanged) under left–right interchanges. However, in 1956, Tsung-Dao Lee and Chen-Ning Yang conjectured[2] that the weak forces responsible for nuclear decays and neutrino interactions might violate parity. Subsequent experiments not only confirmed that feature, but also found that parity violation was maximal: only left-handed particles experienced the weak interaction; right-handed particles were not affected by the weak forces that were known then. Antiparticles, such as antielectrons and antiquarks, exhibited the opposite preference — only their right-handed components participated in weak interactions. For the revolutionary idea of parity violation, Lee and Yang received the physics Nobel prize in 1957.

Beyond parity violation, small differences between the weak interactions of left-handed particles and those of right-handed antiparticles, known as CP violation or matter–antimatter asymmetry, were subsequently observed[3]. Today, some as yet undiscovered

form of CP violation is thought to be responsible for the dominance of matter over antimatter throughout the Universe — a feature responsible for our very existence. Symmetry violation can, indeed, have profound consequences.

Apart from parity violation, electromagnetic and weak interactions are quite similar. Both can be viewed as exchanges of packets (quanta) of energy called bosons. Electromagnetism is mediated by massless photons, whereas heavy, charged W bosons mediate weak interactions. Although some sort of electroweak unification, jointly describing both interactions, seemed natural[4], parity violation caused problems. In 1961, it was shown[5] that unification was possible if, in addition to charged W bosons, another heavy neutral boson, now called the Z boson, also existed. Unfortunately, even then, parity violation made it difficult to accommodate or relate elementary-particle masses. The problem was solved in 1967, when it was demonstrated[6] how the introduction of symmetry breaking through the Higgs mechanism could be used to provide mass. A predicted remnant of that mechanism — the Higgs boson — was detected in 2012 at CERN, Europe's high-energy physics laboratory near Geneva, Switzerland, and François Englert and Peter Higgs were awarded last year's Nobel Prize in Physics for the theoretical work on the Higgs mechanism.

In the early 1970s, support for the existence of the Z boson was observed in neutrino-scattering experiments[7]. But follow-up studies proved inconclusive, in that they did not confirm the parity-violating predictions of electroweak unification. Then an experiment[8,9] called E122, conducted at the SLAC National Accelerator Laboratory in Menlo Park, California, measured a small parity-violating difference between the scattering of right- and left-handed electrons on up and down quarks in a target of deuterium atoms. The up and down quarks are the lightest of the six possible types of quark, and make up all nuclei. This result unequivocally confirmed the parity-violating predictions of electroweak unification. For their work on electroweak unification and its implications, Sheldon Lee Glashow[5], Abdus Salam[10] and Steven Weinberg[6] received the physics Nobel prize in 1979.

During the 35 years since E122 was completed, better sources of right- and left-handed electrons have been developed, experimental techniques have improved and more-intense electron beams have become available. Parity violation has been used for the precise measurement of parameters that describe the electroweak interaction and to investigate nuclear properties. But the parity-violating difference measured in the E122 experiment has not been improved on — until now.

In their study, the Jefferson Lab team decided to redo the SLAC E122 experiment. The researchers worked at lower energy but with much higher intensity and polarization (degree of handedness). As a result, they improved on some aspects of parity-violating differences between the scattering of right- and left-handed electrons on up and down quarks by about a factor of five. With their higher statistics, they were able to untangle the two parity-violating effects: the dominant effect due to electron parity violation, which had already been clearly measured in E122, and a much smaller parity-violating effect attributable to the quarks in the deuterium nuclei, which was beyond the sensitivity of the SLAC experiment.

Why measure such small effects, and so precisely? Perhaps, like mountain-climbing enthusiasts, physicists study them because they are there and represent challenges. However, unlike mountains, in the case of parity-violating effects sometimes smaller is better. Testing the tiny quark parity-violation prediction is a nice example: a deviation from expectations could signal the presence of a new tiny effect. Indeed, the team's measurement probes some types of additional parity-violating effects that could be as much as 30 times weaker than ordinary weak forces. Precision studies also provide access to small nuclear effects that are hard to probe in other ways. An example is the breaking of charge symmetry (the interchange of up and down quarks in deuterium).

Parity-violating polarized electron scattering experiments are expected to continue at the Jefferson Lab, using higher-energy electrons and better particle-detection systems, after upgrades to the facility, now in progress, are completed. One can anticipate better measurements of electroweak parameters, more-refined nuclear-physics studies and improved searches for new interactions.

A great accomplishment can lead to the demise of a scientific endeavour. A good example is the race to put a man on the Moon. That goal started more than 50 years ago and was a spectacular success, but further undertakings ended after the mission was accomplished. Fortunately, electron-scattering studies of parity violation did not suffer that fate. Following the success of E122 at SLAC, the programme changed direction, but improvements in technical expertise and accelerator facilities continued. The Jefferson Lab has taken leadership in polarized-electron scattering initiatives. As long as these initiatives address frontier questions and interesting goals, they should prosper and grow. ■

William J. Marciano *is at the Brookhaven National Laboratory, Upton, New York 11973, USA.*
e-mail: marciano@quark.phy.bnl.gov

1. The Jefferson Lab PVDIS Collaboration *Nature* **506,** 67–69 (2014).
2. Lee, T. D. & Yang, C. N. *Phys. Rev.* **104,** 254–258 (1956).
3. Christenson, J. H., Cronin, J. W., Fitch, V. L. & Turlay, R. *Phys. Rev. Lett.* **13,** 138–140 (1964).
4. Schwinger, J. *Ann. Phys.* **2,** 407–434 (1957).
5. Glashow, S. L. *Nucl. Phys.* **22,** 579–588 (1961).
6. Weinberg, S. *Phys. Rev. Lett.* **19,** 1264–1266 (1967).
7. Hasert, F. J. *et al. Phys. Lett. B* **46,** 138–140 (1973).
8. Prescott, C. Y. *et al. Phys. Lett. B* **77,** 347–352 (1978).
9. Prescott, C. Y. *et al. Phys. Lett. B* **84,** 524–528 (1979).
10. Salam, A. *Conf. Proc. C680519*, 367–377 (1968).

ECOLOGY

Plant diversity rooted in pathogens

Ecologists have long pondered how so many species of plant can coexist locally in tropical forests. It seems that fungal pathogens have a central role, by disadvantaging species where they are locally common. **See Letter p.85**

HELENE C. MULLER-LANDAU

Tropical forests routinely contain more than 200 tree species in a single hectare (Fig. 1). Why don't a few species come to dominate, by chance or by virtue of being better competitors? Multiple hypotheses have been proposed to answer this question, most of which invoke some sort of niche differentiation with respect to resources and/or natural enemies. But despite decades of research, the issue remains unresolved. In this issue, Bagchi *et al.*[1] (page 85) report the results of an elegant field study that clearly implicates natural enemies, specifically fungal pathogens, as crucial to maintaining tropical-plant diversity.

In 1970, ecologists Daniel Janzen[2] and Joseph Connell[3] proposed that natural enemies that target specific host plants maintain high tropical-plant diversity by elevating the mortality of each plant species in areas where it is abundant. Fundamentally, the idea is that host-specialized enemies, including pathogens and insect herbivores, can attack more efficiently and do more damage where their hosts are more plentiful. As a result, each host species fares better when it is

rare and less well as it becomes more common — a phenomenon known as negative density dependence. Many empirical studies have found such negative density dependence in tropical forests[4,5], and the Janzen–Connell hypothesis is the most often cited explanation for these patterns and for high local diversity of plant species in tropical forests. However, niche differences in resource requirements or other factors could also cause negative-density-dependent patterns[6], and few studies have explicitly linked such patterns to particular natural enemies (although see ref. 7 for an exception).

Bagchi *et al.* tested this hypothesis experimentally by using pesticides to remove (or at least reduce) fungal pathogens and, separately, insects at the seedling-establishment stage. Working in a tropical forest in Belize, the authors censused seeds falling into seed traps and seedlings that became established in neighbouring 1-square-metre plots that were treated with a fungicide or with an insecticide, or not treated. In untreated plots, seedling establishment was negatively density dependent and there was a large increase in local species diversity from the seed to the seedling stage, consistent with previous work[4]. Bagchi and colleagues' crucial findings were that fungicide application resulted in the near disappearance of negative density dependence and a drop in seedling species diversity. By contrast, insecticide application merely weakened negative density dependence and led to no change in species diversity, although it did increase the total number of seedlings and caused a dramatic shift in species composition.

Figure 1 | Shades of green. The forest canopy on Barro Colorado Island, Panama, provides visual evidence of how small areas can contain many different tropical tree species. Bagchi and colleagues' findings[1] suggest that fungal pathogens play a crucial part in maintaining this diversity.

This is the first study to explicitly link a particular group of natural enemies to negative density dependence and the maintenance of species diversity in tropical forest plants. It clearly implicates fungal pathogens as the most important drivers of these patterns at the seedling-establishment stage. In the past, there have been more studies of insects than of pathogens as agents of the Janzen–Connell effect — no doubt owing in large part to the greater ease of working with insects. Although insect attack has been found to increase with host-plant density in several tropical plant species[8], the ability of insects to respond to high host density, and thus induce negative density dependence, may ultimately be restricted by their own enemies, such as parasites or predators[9]. Pathogens seem less likely to be similarly checked, which may explain their greater contribution to negative density dependence.

Bagchi and colleagues' results demonstrate that fungal pathogens and insect herbivores influence tropical plant communities in qualitatively different ways. Their distinct roles relate to the two ways in which differences among plant responses to natural enemies can affect species diversity and composition. First, as discussed above, differences in natural enemies can contribute to niche differences that stabilize individual species' abundances and species diversity. Alternatively, or in addition, they can alter differences in competitive ability (fitness) among species[10], thereby modifying species abundances, and potentially which species can successfully compete at all. The large shifts in species composition seen during insecticide treatment suggest that insects have major impacts on fitness differences in this ecosystem. Overall, it seems that fungal pathogens are more important determinants of niche differences and thus species diversity, whereas insects have greater influence on fitness differences and thus species composition.

This work also contributes novel observations on the strength of negative density dependence in different species. Contrasting hypotheses predict either a negative relationship between a species' average abundance and its negative density dependence if greater abundance makes a species more apparent to its enemies, or a positive one if the causality is reversed and lower negative density dependence leads to increased abundance[5]. Previous studies[5,11] that quantified tree abundance over large areas have found that more-abundant species experience less negative density dependence. Bagchi *et al.* find that species that are more abundant as seeds suffer stronger negative density dependence, which at first seems to contradict these earlier findings. However, seed abundance depends not only on tree abundance but also on seed size, which varies widely among tropical tree species. Small-seeded species are likely to be particularly vulnerable to natural enemies, and small seeds are produced in greater numbers, and thus differences in seed size may reconcile Bagchi and colleagues' results with previous work. Future studies should seek to disentangle the roles of species traits and abundances in driving interspecific variation in negative density dependence.

Indeed, this groundbreaking experimental work lays the foundation for a host of studies exploring the roles of natural enemies in structuring tropical-plant diversity. Bagchi *et al.* investigated effects on seedling establishment, a single life stage — integration of such effects over the entire life cycle will ultimately provide a more complete picture. Replication of these experiments across climatic gradients could also test the idea that some climates are more conducive to natural enemies, and that this contributes to greater species diversity of forests in these areas. Furthermore, such comparative studies or others that explicitly manipulate temperature, rainfall or atmospheric carbon dioxide could address how global change affects interactions between plants and natural enemies and thereby illuminate the future of tropical plant diversity. ■

Helene C. Muller-Landau *is at the Smithsonian Tropical Research Institute, Apartado Postal 0843-03092, Panama City, Panama.*
e-mail: mullerh@si.edu

1. Bagchi, R. *et al. Nature* **506,** 85–88 (2014).
2. Janzen, D. H. *Am. Nat.* **104,** 501–528 (1970).
3. Connell, J. H. in *Dynamics of Populations: Proc. Adv. Study Inst. Dynamics of Numbers in Populations*, Oosterbeek (eds den Boer, P. J. & Gradwell, G. R.) 298–312 (Centre Agric. Publ. Docum., 1971).
4. Harms, K. E., Wright, S. J., Calderón, O., Hernández, A. & Herre, E. A. *Nature* **404,** 493–495 (2000).
5. Comita, L. S., Muller-Landau, H. C., Aguilar, S. & Hubbell, S. P. *Science* **329,** 330–332 (2010).
6. Chesson, P. *Annu. Rev. Ecol. Systematics* **31,** 343–366 (2000).
7. Bell, T., Freckleton, R. P. & Lewis, O. T. *Ecol. Lett.* **9,** 569–574 (2006).
8. Hammond, D. S. & Brown, V. K. in *Dynamics of Tropical Communities* (eds Newbery, D. M., Prins, H. H. T. & Brown, N. D.) 51–78 (Blackwell Science, 1998).
9. Visser, M. D., Muller-Landau, H. C., Wright, S. J., Rutten, G. & Jansen, P. A. *Ecol. Lett.* **14,** 1093–1100 (2011).
10. Adler, P. B., Hille Ris Lambers, J. & Levine, J. M. *Ecol. Lett.* **10,** 95–104 (2007).
11. Mangan, S. A. *et al. Nature* **466,** 752–755 (2010).

This article was published online on 22 January 2014.

doi:10.1038/nature12921

Fifty thousand years of Arctic vegetation and megafaunal diet

Eske Willerslev[1]*, John Davison[2]*, Mari Moora[2]*, Martin Zobel[2]*, Eric Coissac[3]*, Mary E. Edwards[4]*, Eline D. Lorenzen[1,5]*, Mette Vestergård[1]*, Galina Gussarova[6,7]*, James Haile[1,8]*, Joseph Craine[9], Ludovic Gielly[3], Sanne Boessenkool[6]†, Laura S. Epp[6]†, Peter B. Pearman[10], Rachid Cheddadi[11], David Murray[12], Kari Anne Bråthen[13], Nigel Yoccoz[13], Heather Binney[4], Corinne Cruaud[14], Patrick Wincker[14], Tomasz Goslar[15,16], Inger Greve Alsos[17], Eva Bellemain[6]†, Anne Krag Brysting[18], Reidar Elven[6], Jørn Henrik Sønstebø[6], Julian Murton[19], Andrei Sher[20]‡, Morten Rasmussen[1], Regin Rønn[21], Tobias Mourier[1], Alan Cooper[22], Jeremy Austin[22], Per Möller[23], Duane Froese[24], Grant Zazula[25], François Pompanon[3], Delphine Rioux[3], Vincent Niderkorn[26], Alexei Tikhonov[27], Grigoriy Savvinov[28], Richard G. Roberts[29], Ross D. E. MacPhee[30], M. Thomas P. Gilbert[1], Kurt H. Kjær[1], Ludovic Orlando[1], Christian Brochmann[6]* & Pierre Taberlet[3]*

Although it is generally agreed that the Arctic flora is among the youngest and least diverse on Earth, the processes that shaped it are poorly understood. Here we present 50 thousand years (kyr) of Arctic vegetation history, derived from the first large-scale ancient DNA metabarcoding study of circumpolar plant diversity. For this interval we also explore nematode diversity as a proxy for modelling vegetation cover and soil quality, and diets of herbivorous megafaunal mammals, many of which became extinct around 10 kyr BP (before present). For much of the period investigated, Arctic vegetation consisted of dry steppe-tundra dominated by forbs (non-graminoid herbaceous vascular plants). During the Last Glacial Maximum (25–15 kyr BP), diversity declined markedly, although forbs remained dominant. Much changed after 10 kyr BP, with the appearance of moist tundra dominated by woody plants and graminoids. Our analyses indicate that both graminoids and forbs would have featured in megafaunal diets. As such, our findings question the predominance of a Late Quaternary graminoid-dominated Arctic mammoth steppe.

It can be argued that Arctic vegetation during the proximal Quaternary (the last circa 50 kyr) is less well understood than the ecology and population dynamics of the mammals that consumed it, despite the overall uniformity and low floristic diversity of Arctic vegetation[1,2]. Analyses of vegetation changes during this interval have been based mainly on fossil pollen. Although highly informative, records tend to be biased towards high pollen producers such as many graminoids (grasses, sedges and rushes) and *Artemisia*, which can obscure the abundance of other forms such as many insect-pollinated forbs[1]. Arctic pollen records are rarely comprehensively identified to species level, which underestimates actual diversity[3]. These problems are to some extent ameliorated by plant macrofossil studies (for example, ref. 4), which may provide detailed records of local vegetation. However, macrofossil studies are far less common, have their own taxonomic constraints, and usually cannot provide quantitative estimates of abundance.

In recent years, a complementary approach has emerged that uses plant and animal ancient DNA preserved in permafrost sediments[5]. Such environmental DNA[6] does not derive primarily from pollen, bones or teeth, but likely from above- and below-ground plant biomass, faeces, discarded cells and urine preserved in sediments[7–9]. Like macrofossils, environmental DNA appears to be local in origin[6,10–12] and, in principle, the survival of a few fragmented DNA molecules is sufficient for retrieval and taxonomic identification[13].

Environmental DNA can supply the fraction of the plant community not readily identifiable by pollen analysis and, to some extent, macrofossils, particularly in vegetation dominated by non-woody growth forms[7].

[1]Centre for GeoGenetics, Natural History Museum, University of Copenhagen, Oster Voldgade 5-7, DK-1350 Copenhagen K, Denmark. [2]Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, 40 Lai Street, 51005 Tartu, Estonia. [3]Laboratoire d'Ecologie Alpine (LECA) CNRS UMR 5553, University Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France. [4]Geography and Environment, University of Southampton, Southampton SO17 1BJ, UK. [5]Department of Integrative Biology, University of California Berkeley, 1005 Valley Life Sciences Building, Berkeley, 94720 California, USA. [6]National Centre for Biosystematics, Natural History Museum, University of Oslo, PO Box 1172, Blindern, NO-0318 Oslo, Norway. [7]Department of Botany, Saint Petersburg State University, Universitetskaya nab. 7/9, 199034 Saint Petersburg, Russia. [8]Ancient DNA Laboratory, Veterinary and Life Sciences School, Murdoch University, 90 South Street, Perth, 6150 Western Australia, Australia. [9]Division of Biology, Kansas State University, Manhattan, 66506-4901 Kansas, USA. [10]Landscape Dynamics Unit, Swiss Federal Research Institute WSL, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland. [11]Institut des Sciences de l'Evolution de Montpellier, UMR 5554 Université Montpellier 2, Bat.22, CC061, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France. [12]University of Alaska Museum of the North, Fairbanks, 99775-6960 Alaska, USA. [13]Department of Arctic and Marine Biology, UiT, The Arctic University of Norway, NO-9037 Tromsø, Norway. [14]Genoscope, Institut de Genomique du Commissariat à l'Energie Atomique (CEA), 91000 Evry, France. [15]Adam Mickiewicz University, Faculty of Physics, Umultowska 85, 61-614 Poznań, Poland. [16]Poznań Radiocarbon Laboratory, Poznań Science and Technology Park, Rubież 46, 61-612 Poznań, Poland. [17]Tromsø University Museum, NO-9037 Tromsø, Norway. [18]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, P.O. Box 1066, Blindern, NO-0316 Oslo, Norway. [19]Permafrost Laboratory, Department of Geography, University of Sussex, Brighton BN1 9QJ, UK. [20]Institute of Ecology and Evolution, Russian Academy of Sciences, 33 Leninsky Prospect, 119071 Moscow, Russia. [21]Department of Biology, Terrestrial Ecology, Universitetsparken 15, DK- 2100 Copenhagen Ø, Denmark. [22]Australian Centre for Ancient DNA, School of Earth & Environmental Sciences, University of Adelaide, Adelaide, 5005 South Australia, Australia. [23]Department of Geology/Quaternary Sciences, Lund University Sölvegatan 12, SE-223 62 Lund, Sweden. [24]Department of Earth and Atmospheric Sciences, University of Alberta, T6G 2E3 Edmonton, Alberta, Canada. [25]Government of Yukon, Department of Tourism and Culture, Yukon Palaeontology Program, PO Box 2703 L2A, Y1A 2C6 Whitehorse, Yukon Territory, Canada. [26]INRA, UMR1213 Herbivores, F-63122 Saint-Genès-Champanelle, France. [27]Zoological Institute of Russian Academy of Sciences, Universitetskaya nab. 1, 199034 Saint-Petersburg, Russia. [28]Institute of Applied Ecology of the North of North-Eastern Federal University, Belinskogo Street 58, 677000 Yakutsk, Republic of Sakha (Yakutia), Russia. [29]Centre for Archaeological Science, School of Earth and Environmental Sciences, University of Wollongong, Wollongong, 2522 New South Wales, Australia. [30]Division of Vertebrate Zoology/Mammalogy, American Museum of Natural History, New York, 10024 New York, USA. †Present addresses: Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, PO Box 1066, Blindern, NO-0318 Oslo, Norway (S.B.); Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Research Unit Potsdam, Telegrafenberg A 43, 14473 Potsdam, Germany (L.S.E.); SpyGen, Savoie Technolac, 17 allée du lac Saint André, BP 274, 73375 Le Bourget-du-Lac Cedex, France (E.B.).
*These authors contributed equally to this work.
‡Deceased

For most plant groups, DNA permits identification at lower taxonomic levels than pollen[14]. In addition, environmental DNA records have proven to reflect not only the qualitative but also the quantitative diversity of above-ground plant[12] and animal taxa[9], as determined from modern subsurface soils.

Leaching of DNA through successive stratigraphic zones may be an issue in temperate conditions[9,11] but not in permafrost[6] or in sediments that have only recently thawed[15]. Re-deposition of sediments and organics can confound results, which is also the case for pollen and macrofossils[7,16], but can be avoided and accounted for by careful site selection and by excluding rare DNA sequence reads[16]. For Quaternary permafrost settings, at least, taphonomic bias due to differences in DNA survival across plant groups does not appear to be of concern (see Methods section 4.0 on taphonomy), as has been shown by a comparative permafrost ancient DNA study of plants and their associated fungi[8].

Reconstruction of Arctic vegetation from permafrost

We collected 242 sediment samples from 21 sites across the Arctic (Fig. 1 and Extended Data Table 1). Ages were determined by accelerator mass spectrometry radiocarbon (^{14}C) dating, and are reported here in thousands of calibrated (calendar) years BP (Extended Data Fig. 1 and Supplementary Data 1). We sequenced the short P6 loop sequence of the *trn*L plastid (gene encoding chloroplast transfer RNA for leucine) region and a part of the ITS1 spacer region through metabarcoding (Methods section 3.0), generating a total of 14,601,839 *trn*L plant DNA sequence reads and 1,652,857 internal transcribed spacer (ITS) reads. Reads were identified by comparison with (1) the Arctic *trn*L taxonomic reference library[14], which we extended with ITS sequences for three families; (2) a new north boreal *trn*L taxonomic reference library constructed by sequencing 1,332 modern plant samples representing 835 species; and (3) GenBank, using the program ecoTag (Supplementary Data 2 and Methods section 3.0). Basic statistics, *in silico* analyses, and additional experiments were carried out to check data reliability (Extended Data Fig. 2 and Extended Data Table 2). We grouped the identified molecular operational taxonomic units (MOTUs) into three distinct intervals (Fig. 2a): (1) pre-Last Glacial Maximum (LGM) (50–25 kyr BP), a period of fluctuating climate; (2) LGM (25–15 kyr BP), a period of constantly cold and dry conditions; and (3) post-LGM (15–0 kyr BP), which includes the current interglacial, characterized by relatively higher temperatures[17].

Shifts in plant community composition

To address compositional changes in vegetation across space and time we used a generalized linear model and permutational multivariate analysis of variance (PERMANOVA) (Supplementary Data 3 and Methods section 6.0). We find that (1) the composition of plant MOTU assemblages differed significantly across the three intervals (pseudo-$F = 6.77$, $P < 0.001$, Extended Data Fig. 3a–e), with pre-LGM and post-LGM plant assemblages differing the most (Extended Data Fig. 3f); (2) the greater the spatial distance separating a pair of samples within each time period, the less similar their composition ($P < 0.001$); and (3) LGM assemblages were the most homogeneous across space and post-LGM assemblages were the most heterogeneous (Fig. 2).

LGM pollen spectra show high floristic richness compared to other intervals (for example, ref. 1). This is due to the limited occurrence of woody taxa with high pollen production, which in turn proportionately emphasizes less-productive taxa. By contrast, our DNA data reveal that plant diversity was lowest during LGM relative to other intervals (Fig. 2a). Plant assemblages became more similar to each other and the estimated number of MOTUs decreased from pre-LGM to LGM (Fig. 2a), with many taxa absent that had previously been well represented

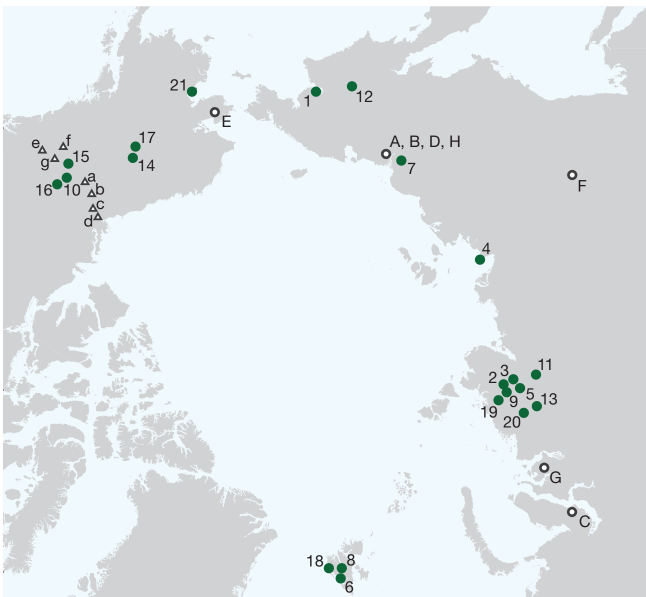


Figure 1 | Sample localities. A total of 242 permafrost samples were collected from 21 sites, shown by green dots (1–21). Eight ancient megafauna gut and coprolite samples (A–H) are shown by grey hollow circles, and seven modern nematode localities are shown by grey hollow triangles (a–g). (1) Anadyr, (2) Baskura Peninsula, (3) Bol'shaya Balakhnaya, (4) Buor Khaya, (5) Cape Sabler, (6) Colesdalen, (7) Duvanny Yar, (8) Endalen, (9) Federov Island, (10) Goldbottom, (11) Khatanga, (12) Maine River, (13) Ovrazhny Peninsula, (14) Purgatory, (15) Quartz Creek, (16) Ross Mine, (17) Stevens Village, (18) Stuphallet, (19) Taimyr Lake, (20) Upper Taymyr River, (21) Zagoskin Lake, (A) Drevniy Creek Mammoth, (B) Bison, (C) Lyuba Mammoth, (D) Kolyma Rhino, (E) Last Chance Creek Horse, (F) Churapcha Rhino, (G) Mongochen Mammoth, (H) Finish Creek Valley Mammoth, (a) Blackstone River, (b) Ogilvie Mountains, (c) Eagle Plains South, (d) Eagle Plains North, (e) Little Atlin Lake, (f) Kluane Lake, (g) Carmacks.

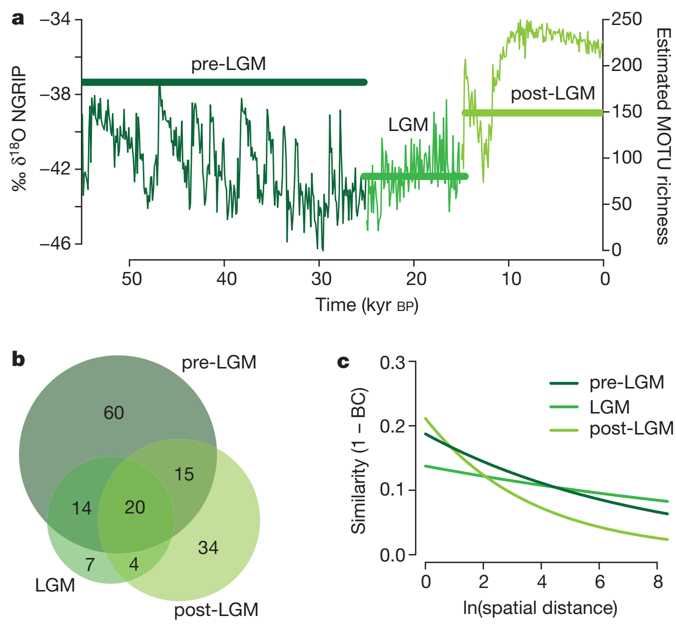


Figure 2 | Taxonomic diversity of Arctic plant assemblages during the last 50 kyr. Taxon composition was estimated by high-throughput sequencing of DNA from 242 permafrost samples. A total of 154 MOTUs were detected. **a**, Index of ambient temperature (continuous line; oxygen isotope concentration, North Greenland Ice Core Project, NGRIP[50]) and estimated MOTU number (horizontal bars; second-order jackknife) are shown for three palaeoclimatic periods: pre-LGM (>25 kyr BP, $n = 149$), LGM (25–15 kyr BP, $n = 32$) and post-LGM (<15 kyr BP, $n = 61$). **b**, MOTU counts recorded uniquely in each palaeoclimatic period and shared among periods. **c**, Modelled decline in similarity (1 − Bray–Curtis (BC) dissimilarity) between pairs of plant assemblages from the same palaeoclimatic period in relation to the spatial distance separating them.

(Fig. 2b). In addition, although the LGM flora was largely a subset of the pre-LGM flora, the post-LGM flora was different (Fig. 2b), with pronounced geographic differentiation (Fig. 2c).

Steppe-tundra

Owing to the low taxonomic resolution of previously published vegetation reconstructions, it remains undetermined whether Arctic vegetation during the last part of the Quaternary was a form of tundra or more like steppe (for example, refs 18, 19). Small-scale contemporary analogues range from low-productivity fellfields and cryoxeric steppe communities to more productive dry Arctic steppe-to-tundra gradients. Our sediment DNA plant sequence data from ~50–12 kyr BP encompass taxa that typify both tundra and Arctic steppe environments. These include taxa that are today typical of dry and/or disturbed sites (for example, *Bromus pumpellianus*, *Artemisia frigida*, *Plantago canescens*, *Anemone patens*), saline soils (*Puccinellia*, *Armeria*), moist habitats (*Caltha*) and rocky or fellfield habitats (*Dryas*, *Draba*), plus a woody component dominated by *Salix* (Supplementary Data 4 and 5). A spatial and/or temporal mosaic of plant communities is indicated (Methods section 6.0), as is seen in floristically rich macrofossil records[4]. The most common MOTU in the pre-LGM and LGM samples is Anthemideae group 1 (*Artemisia*, *Achillea*, *Chrysanthemum*, *Tanacetum*), which underscores the importance in regional pollen assemblages of Asteraceae in general and *Artemisia* in particular[1]. *Equisetum* and *Eriophorum* are important only in postglacial assemblages, reflecting moister soil conditions. Increases in aquatic taxa (Supplementary Data 4 and 5) also indicate a predominance of moister substrates in the later part of the post-LGM period. These findings indicate a shift from dry steppe-tundra to moist tundra in the early part of the post-LGM period—a change widely reported in other proxy studies.

Nematode assemblage composition is known to change with vegetation cover[20], moisture[21] and organic resource inputs[22]. Therefore, to obtain a complementary proxy for vegetation cover and soil quality, we characterized the soil nematode fauna of contemporary mesic shrub tundra and subarctic steppe on well-drained loess soils in Yukon Territory, Canada (Fig. 1 and Extended Data Table 3). The relative proportion of the nematode families Teratocephalidae and Cephalobidae varied among vegetation types ($P < 0.001$, nested ANOVA), and indicator species analysis[23] confirmed that Teratocephalidae (indicator value = 0.98, $P = 0.001$) and Cephalobidae (indicator value = 0.98, $P = 0.001$) are very good indicators of tundra and steppe vegetation, respectively (Fig. 3).

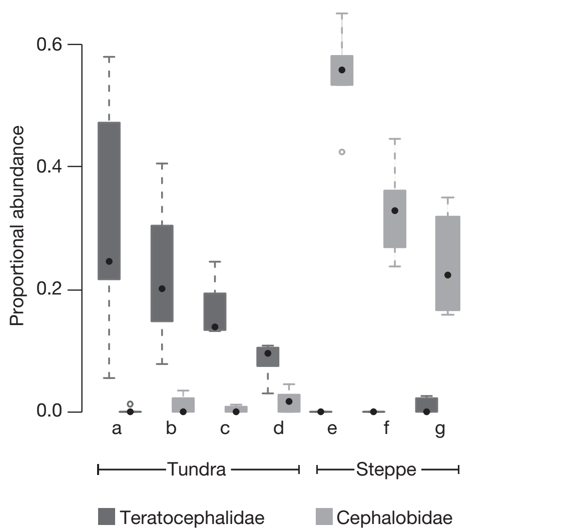


Figure 3 | Proportional abundance of two families—Teratocephalidae and Cephalobidae—among the total soil nematode community at contemporary tundra and steppe sites in Yukon, Canada. Teratocephalidae, dark; Cephalobidae, light. Letters a–g correspond to sample localities (Fig. 1). Median (central dot), quartile (box), maximum and minimum (whiskers) and outlying values (points) are shown.

These findings are in agreement with previous studies restricted to subarctic Sweden[24,25] and alpine and subalpine habitats[26,27]. We amplified short DNA sequences from these two taxa from 17 sediment samples analysed for plant DNA from Yukon and northeastern Siberia. We detected Cephalobidae DNA in almost all samples, whereas Teratocephalidae was detected at a higher frequency in samples younger than 10 kyr BP than in the pre-LGM and LGM samples (Extended Data Table 4). These results support our inferences from plant sequence data and indicate a transition from relatively dry tundra and steppe towards more moist tundra during the post-LGM interval.

Forb dominance and megafaunal diets

To assess structural and functional shifts in the plant assemblages, we investigated temporal changes in the relative abundance of different growth forms. Our DNA results show that pre-LGM vegetation was dominated by forbs, the relative share of which increased during the LGM, whereas graminoids constituted less than 20% of the total read count (Fig. 4a). These results persisted when we corrected for observed modern representational bias[12] (Methods sections 4.0 and 5.3).

Continued forb dominance during the LGM implies that similar proportions of forbs and graminoids were maintained through this period, despite the significant decline in floristic diversity (Fig. 2a, b). Our findings contrast with pollen-based reconstructions, which have emphasized dominance of graminoids in the unglaciated Arctic and adjacent regions, particularly during the LGM, and are exemplified by the widely used term mammoth steppe[19]. Rather, our results show that vegetation was forb-dominated in both overall abundance of MOTUs and in floristic richness (Fig. 4a, b and Extended Data Fig. 3g, h), in agreement with macrofossil data that show a diversity of forbs of mixed ecological preference (for example ref. 4).

We explored whether forbs were prominent in habitats favoured by megafauna by analysing 25 dated (47–20 kyr BP) sediment samples from Main River, Siberia, using *trn*L plastid plant and 16S mitochondrial DNA mammal primers. We found that the mean proportion of forbs was higher in samples from which herbivorous megafaunal DNA had been retrieved ($n = 18$; for example, woolly mammoth, woolly rhinoceros, horse, reindeer and elk) than in samples lacking such DNA ($n = 7$; Fig. 4c and Extended Data Table 5). Although suggestive of co-occurrence of megafauna in forb-dominated settings, these results should be regarded as tentative, and further studies are needed to verify if this is indeed a general trend.

We also investigated whether megafaunal diets revealed the level of forb dominance observed in permafrost sediment samples. Using standardized methods, we genetically characterized intestinal/stomach contents and coprolites recovered from eight specimens of woolly mammoth, woolly rhinoceros, bison and horse from Siberia and Alaska, dated >55–21 kyr BP (Extended Data Table 6 and Methods sections 3.0 and 7.3). Although ingested plant remains are often difficult to identify morphologically, they can be accurately identified[28,29] and roughly quantified[30] using DNA. The majority of these samples are dominated by forbs, which comprise 0.63 ± 0.12 of the sequences, compared to 0.27 ± 0.16 expressing graminoid sequences (Fig. 4d and Supplementary Data 6). These results suggest that megafaunal species supplemented their diets with high-protein forbs rather than specializing more or less exclusively on grasses.

To confirm the reliability of our *trn*L approach for estimating herbivore diet, we analysed 50 rumen samples of sheep-feed diets with varying proportions of forbs (white clover (*Trifolium repens*)) and graminoids (ryegrass (*Lolium perenne*)) (Methods section 5.4). As seen in Fig. 4e, the Pearson correlation coefficient between the actual fraction of forbs in these diets and the proportion of forbs estimated with the DNA-based approach was highly significant ($r^2 = 0.75$, $P < 10^{-15}$).

Discussion

Our observations of high forb abundance in the Terminal Pleistocene may merely reflect vegetation response to glacial climates, but there are

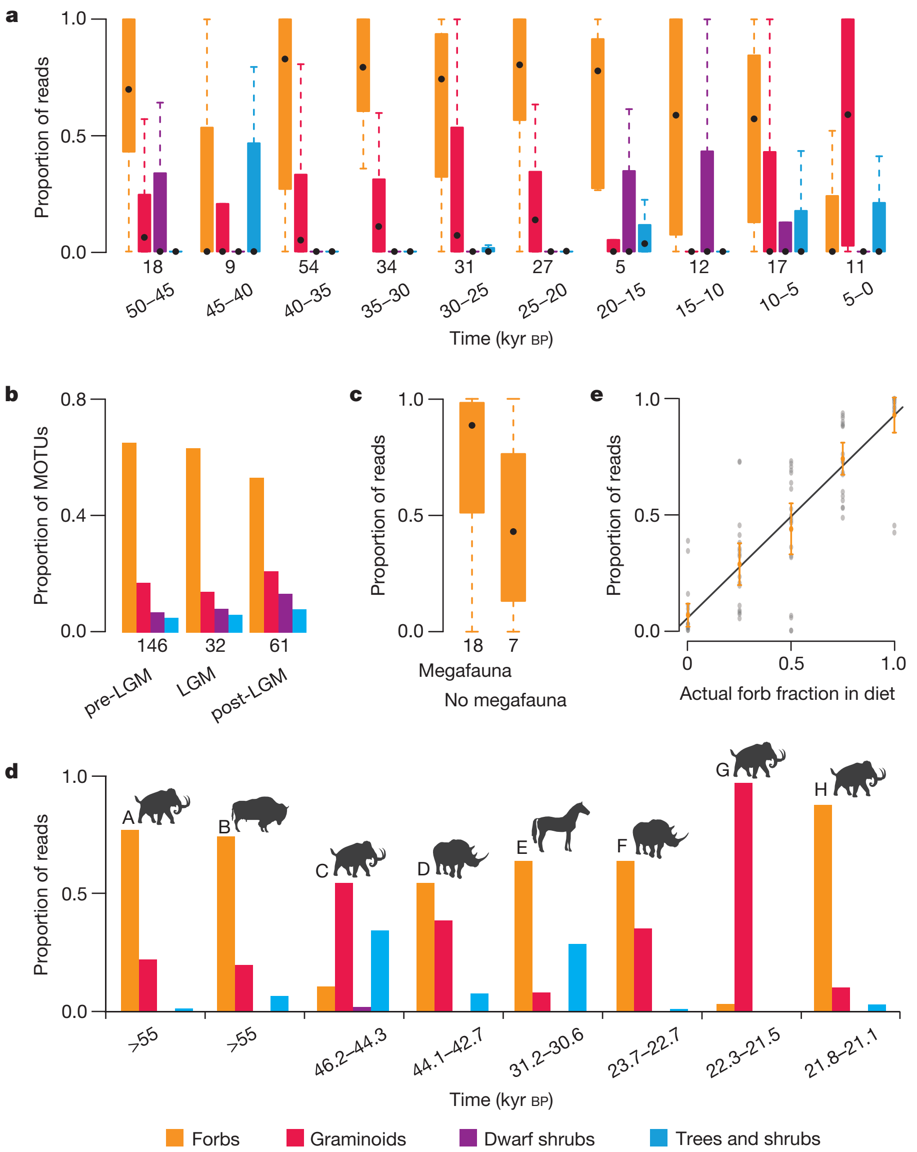


Figure 4 | Plant growth form composition over time and across sample types, estimated by high-throughput sequencing of DNA from 242 permafrost samples. **a**, Proportions of DNA reads corresponding to taxa exhibiting different growth forms, binned over 5 kyr time intervals. The analysis included all sediment samples except 21 Svalbard samples and three further samples for which no growth form information was available. **b**, Number of MOTUs exhibiting different growth forms as a proportion of total MOTU richness in all informative samples for each palaeoclimatic period. **c**, The proportional abundance of forbs in samples from Main River, Siberia (dated 47,100–19,850 yr BP) where megafauna were or were not detected. **d**, Proportions of DNA reads corresponding to different growth forms in megafauna diet, determined from analysis of eight gut and coprolite samples from late Quaternary megafauna species (woolly mammoth, woolly rhinoceros, bison and horse). Letters A–H correspond to the individual samples (Fig. 1). The 95.4% calibrated age range of each sample is shown; '> 55' indicates that the sample was too old to provide a finite radiocarbon age. **e**, Reliability of the *trn*L approach for estimating forb and graminoid abundance in diet analyses. Sheep were fed with known amounts of forbs (*Trifolium repens*) and graminoids (*Lolium perenne*), and the rumen content analysed using the same DNA-based approach as implemented above. Grey dots are raw data points, orange dots and lines represent the means and ± standard errors for diets containing different fractions of forbs. The grey line is a linear model fit. Numbers immediately below the columns in **a**, **b** and **c** indicate sample sizes. Median (central dot), quartile (box), maximum and minimum (whiskers) values are shown in **a** and **c**.

other possibilities[1]. An abundant megafauna would have caused significant trampling[31], enhancing gap-based recruitment[32], which could favour forbs[33]. Coupled with nitrogen input from wide-ranging herbivores[34], forbs may out-compete grasses[35]. Furthermore, a diet rich in forbs may help to explain how numerous large animals were sustained; forbs may be more nutrient-rich (for example, ref. 35) and more easily digested[36] than grasses. However, a feedback loop that maintained nutritious and productive forage and supported large mammalian populations in glacial climate regimes may have been impossible to maintain after deglaciation, as C:N ratios increased with global warming[37], and the potential breakdown of the megafauna–forb interaction would have been exacerbated by declining mammalian populations. In contemporary tundra and steppe (the latter often called grasslands), graminoids are generally perceived to be the dominant growth form in large herbivore habitats (for example, refs 38, 39). Our data, which unearth 50 kyr of Arctic vegetation history, call this perception into question.

METHODS SUMMARY

Plant fragments or soil matrix organics were ^{14}C-dated using accelerator mass spectrometry and measured ages were converted into calendar years[40]. Permafrost sampling, DNA extraction, PCR amplification and taxon identification (for example, ref. 41) followed established procedures. Most vascular taxa are covered by ref. 42, and nomenclature is provided accordingly; for the remaining taxa nomenclature follows ref. 43. Dissimilarity between plant assemblages was quantified using pairwise Bray–Curtis distance[44]. Variation in assemblage dissimilarity was decomposed using PERMANOVA[45] and visualized using non-metric multidimensional scaling[46,47]. We used a distance decay approach[48] and a generalized linear model to model variation in plant community assemblages over space and time. Growth form composition of communities was compiled from species trait databases[49]. Differences in the trait composition of assemblages in adjacent climatic periods were compared to a null model assuming random assortment from the previous interval. Nematode faunas of 35 contemporary sediment samples were morphologically determined. Presence of two indicator families (Teratocephalidae for tundra and Cephalobidae for steppe) was genetically determined in 17 ancient sediment samples. Megafaunal DNA and faeces and gut content were determined genetically following established methods. For a detailed discussion, see Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 July; accepted 28 November 2013.

1. Anderson, P. M., Edwards, M. E. & Brubaker, L. B. in *The Quaternary Period in the United States. Developments in Quaternary Science* (eds Gillespie, A. E., Porter, S. C. & Atwater, B. F.) 427–440 (Elsevier, 2003).
2. Murray, D. F. in *Arctic and Alpine Biodiversity: Patterns, Causes and Ecosystem Consequences* (eds Chapin, F. S. & Körner, C.) 21–32 (Springer, 1995).
3. Lamb, H. F. & Edwards, M. E. in *Vegetation History. Handbook of Vegetation Science 7* (eds Huntley, B. & Webb, T. III) 519–555 (Kluwer Academic, 1988).

4. Kienast, F., Schirrmeister, L., Siegert, C. & Tarasov, P. E. Palaeobotanical evidence for warm summers in the East Siberian Arctic during the last cold stage. *Quat. Res.* **63,** 283–300 (2005).
5. Willerslev, E. *et al.* Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* **300,** 791–795 (2003).
6. Haile, J. *et al.* Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl Acad. Sci. USA* **106,** 22352–22357 (2009).
7. Jørgensen, T. *et al.* A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Mol. Ecol.* **21,** 1989–2003 (2012).
8. Lydolph, M. C. *et al.* Beringian paleoecology inferred from permafrost-preserved fungal DNA. *Appl. Environ. Microbiol.* **71,** 1012–1017 (2005).
9. Andersen, K. *et al.* Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Mol. Ecol.* **21,** 1966–1979 (2012).
10. Parducci, L. *et al.* Glacial survival of boreal trees in northern Scandinavia. *Science* **335,** 1083–1086 (2012).
11. Haile, J. *et al.* Ancient DNA chronology within sediment deposits: are paleobiological reconstructions possible and is DNA leaching a factor? *Mol. Biol. Evol.* **24,** 982–989 (2007).
12. Yoccoz, N. G. *et al.* DNA from soil mirrors plant taxonomic and growth form diversity. *Mol. Ecol.* **21,** 3647–3655 (2012).
13. Willerslev, E. & Cooper, A. Ancient DNA. *Proc. R. Soc. Lond. B* **272,** 3–16 (2005).
14. Sønstebø, J. H. *et al.* Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Mol. Ecol. Resour.* **10,** 1009–1018 (2010).
15. Hebsgaard, M. B. *et al.* The farm beneath the sand—an archaeological case study on ancient 'dirt' DNA. *Antiquity* **83,** 430–444 (2009).
16. Arnold, L. J. *et al.* Paper II - Dirt, dates and DNA: OSL and radiocarbon chronologies of perennially frozen sediments in Siberia, and their implications for sedimentary ancient DNA studies. *Boreas* **40,** 417–445 (2011).
17. Hopkins, D. M. in *Paleoecology of Beringia* (eds Hopkins, D. M., Matthews, J. V. Jr, Schweger, C. E. & Young, S. B.) 3–28 (Academic, 1982).
18. Ritchie, J. C. & Cwynar, L. C. in *Paleoecology of Beringia* (eds Hopkins, D. M., Matthews, J. V. Jr, Schweger, C. E., Young, S. B. & Stanley, V.) 113–126 (Academic, 1982).
19. Guthrie, R. D. *Frozen Fauna of the Mammoth Steppe* (Univ. Chicago Press, 1990).
20. Yeates, G. W. Diversity of nematode faunae under three vegetation types on a pallic soil in Otago, New Zealand. *NZ J. Zool.* **23,** 401–407 (1996).
21. Sohlenius, B. Influence of climatic conditions on nematode coexistence — a laboratory experiment with a coniferous forest soil. *Oikos* **44,** 430–438 (1985).
22. Yeates, G. W. Nematodes as soil indicators: functional and biodiversity aspects. *Biol. Fertil. Soils* **37,** 199–210 (2003).
23. Dufrêne, M. & Legendre, P. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol. Monogr.* **67,** 345–366 (1997).
24. Ruess, L., Michelsen, A. & Jonasson, S. Simulated climate change in subarctic soils: responses in nematode species composition and dominance structure. *Nematology* **1,** 513–526 (1999).
25. Sørensen, L. I., Mikola, J., Kytoviita, M.-M. & Olofsson, J. Trampling and spatial heterogeneity explain decomposer abundances in a sub-Arctic grassland subjected to simulated reindeer grazing. *Ecosystems* **12,** 830–842 (2009).
26. Popovici, I. & Ciobanu, M. Diversity and distribution of nematode communities in grasslands from Romania in relation to vegetation and soil characteristics. *Appl. Soil Ecol.* **14,** 27–36 (2000).
27. Hoschitz, M. & Kaufmann, R. Nematode community composition in five alpine habitats. *Nematology* **6,** 737–747 (2004).
28. Poinar, H. N. *et al.* Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis*. *Science* **281,** 402–406 (1998).
29. Hofreiter, M. *et al.* A molecular analysis of ground sloth diet through the last glaciation. *Mol. Ecol.* **9,** 1975–1984 (2000).
30. Soininen, E. M. E. *et al.* Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Front. Zool.* **6,** 16 (2009).
31. Zimov, S. A., Zimov, N. S., Tikhonov, A. N. & Chapin, F. S. I. Mammoth steppe: a high-productivity phenomenon. *Quat. Sci. Rev.* **57,** 26–45 (2012).
32. Owen-Smith, N. Pleistocene extinctions: the pivotal role of megaherbivores. *Paleobiology* **13,** 351–362 (1987).
33. Austrheim, G. & Eriksson, O. Recruitment and life-history traits of sparse plant species in subalpine grasslands. *Can. J. Bot.* **81,** 171–182 (2003).
34. Wardle, D. A. & Bardgett, R. D. Human-induced changes in large herbivorous mammal density: the consequences for decomposers. *Front. Ecol. Environ.* **2,** 145–153 (2004).
35. Güsewell, S. N. P ratios in terrestrial plants: variation and functional significance. *New Phytol.* **164,** 243–266 (2004).
36. Cornelissen, J. *et al.* Leaf digestibility and litter decomposability are related in a wide range of subarctic plant species and types. *Funct. Ecol.* **18,** 779–786 (2004).
37. McLauchlan, K. K., Williams, J. J., Craine, J. M. & Jeffers, E. S. Changes in global nitrogen cycling during the Holocene epoch. *Nature* **495,** 352–355 (2013).
38. van der Wal, R. Do herbivores cause habitat degradation or vegetation state transition? Evidence from the tundra. *Oikos* **114,** 177–186 (2006).
39. Bråthen, K. A. *et al.* Induced shift in ecosystem productivity? Extensive scale effects of abundant large herbivores. *Ecosystems* **10,** 773–789 (2007).
40. Reimer, P. J. *et al.* IntCal09 and Marine09 radiocarbon age calibration curves, 0–50,000 years cal BP. *Radiocarbon* **51,** 1111–1150 (2009).
41. Taberlet, P. *et al.* Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* **35,** e14 (2007).
42. Elven, R., Murray, D. F., Razzhivin, V. Y. & Yurtsev, B. A. *Annotated Checklist of the Panarctic Flora (PAF)* http://nhm2.uio.no/paf/ (Natural History Museum, Univ. Oslo, 2011).
43. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37,** D5–D15 (2009).
44. Bray, J. R. & Curtis, J. T. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27,** 325–349 (1957).
45. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. *Austral. Ecol.* **26,** 32–46 (2001).
46. Shepard, R. N. The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika* **27,** 125–140 (1962).
47. Shepard, R. N. The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika* **27,** 219–246 (1962).
48. Nekola, J. C. & White, P. S. The distance decay of similarity in biogeography and ecology. *J. Biogeogr.* **26,** 867–878 (1999).
49. Klotz, S., Kühn, I. & Durka, W. *BIOLFLOR* (Bundesamt für Naturschutz, 2002).
50. NGRIP dating group, 2008. IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series # 2008-034. NOAA/NCDC Paleoclimatology Program, Boulder CO, USA. (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Lister, R. D. Guthrie, M. Hofreiter and L. Parducci for thoughts and discussions on our findings and K. Andersen for help identifying possible contamination. We thank T. B. Brand, P. S. Olsen, V. Mirré, L. J. Gillespie, J. M. Saarela, J. Doubt, M. Lomonosova, D. Shaulo, J. E. Eriksen, S. Ickert-Bond, T. Ager, D. Bielman, M. Hajibabaei, A. Telka and S. Zimov for help and providing samples. We thank the Danish National Sequencing Centre. This work was supported by the European Union 6th framework project ECOCHANGE (GOCE-2006-036866, coordinated by P.T.), the Danish National Research Foundation (Centre of Excellence to E.W.), the European Regional Development Fund (Centre of Excellence FIBIR and IUT 20-28 to J.D., M.M. and M.Z.), the Research Council of Norway (191627/V40 to C.B.), the Australian Research Council (DP0558446 to R.G.R.), a Marie Curie International Outgoing Fellowship (PIOF-GA-2009-253376 to E.D.L.) and a Carlsberg Foundation Fellowship (to M.V.).

Author Contributions The paper represents the joint efforts of several research groups, headed by various people within each group. Rather than publishing a number of independent papers, we have chosen to combine our data in this paper in the belief that this creates a more comprehensive story. The authorship reflects this joint effort. The ECOCHANGE team designed and initiated the project. E.W., M.E.E., J.M., E.D.L., M.V., G.G., J.H., J.C., I.G.A., P.M., D.F., G.Z., A.T., J.A., A.S., G.S., R.G.R., R.D.E.M., M.T.P.G., A.C. and K.H.K. collected the samples. G.G., R.E., A.K.B., J.H.S., C.B., L.G., E.C. and P.T. constructed the plant DNA taxonomic reference libraries and provided taxonomic assignments of the sediment data with input from I.G.A., E.B., S.B., L.S.E., M.E.E. and D.M. E.D.L., M.V., J.H., L.S.E., S.B., C.C., P.W., L.G., G.G. and J.H.S. conducted the genetics laboratory work. T.G. did the dating. F.P., D.R. and V.N. produced and analysed the data concerning the reliability of the *trn*L approach for estimating herbivore diet. J.D., M.M., M.Z., E.C., M.V., M.R., J.C., S.B., P.B.P., R.C., H.B., R.R., T.M. and P.T. did the analyses. E.D.L. and J.D. produced the figures. E.W. wrote most of the text with input from all authors, in particular J.D., M.M., M.Z., E.D.L., M.E.E., M.V., P.B.P., D.M., K.A.B., N.Y., L.O., C.B., P.T. and R.D.E.M.

Author Information All the raw and filtered data concerning plants, nematodes, megafauna and sheep diet are available either from Extended Data and Supplementary Data, or from the Dryad Digital Repository: http://doi.org/10.5061/dryad.ph8s5. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.W. (ewillerslev@snm.ku.dk).

doi:10.1038/nature12988

In vivo discovery of immunotherapy targets in the tumour microenvironment

Penghui Zhou[1]*, Donald R. Shaffer[1]*†, Diana A. Alvarez Arias[1], Yukoh Nakazaki[1], Wouter Pos[1], Alexis J. Torres[2], Viviana Cremasco[1], Stephanie K. Dougan[3], Glenn S. Cowley[4], Kutlu Elpek[1]†, Jennifer Brogdon[5], John Lamb[6], Shannon J. Turley[1], Hidde L. Ploegh[3], David E. Root[4], J. Christopher Love[2], Glenn Dranoff[1], Nir Hacohen[4], Harvey Cantor[1] & Kai W. Wucherpfennig[1]

Recent clinical trials showed that targeting of inhibitory receptors on T cells induces durable responses in a subset of cancer patients, despite advanced disease. However, the regulatory switches controlling T-cell function in immunosuppressive tumours are not well understood. Here we show that such inhibitory mechanisms can be systematically discovered in the tumour microenvironment. We devised an *in vivo* pooled short hairpin RNA (shRNA) screen in which shRNAs targeting negative regulators became highly enriched in murine tumours by releasing a block on T-cell proliferation upon tumour antigen recognition. Such shRNAs were identified by deep sequencing of the shRNA cassette from T cells infiltrating tumour or control tissues. One of the target genes was *Ppp2r2d*, a regulatory subunit of the PP2A phosphatase family. In tumours, *Ppp2r2d* knockdown inhibited T-cell apoptosis and enhanced T-cell proliferation as well as cytokine production. Key regulators of immune function can therefore be discovered in relevant tissue microenvironments.

Recent work has shown that cytotoxic T cells have a central role in immune-mediated control of cancer[1–7]. T cells are able to specifically detect and eliminate cancer cells following T-cell receptor (TCR)-mediated recognition of tumour-derived peptides bound to MHC proteins[8]. A series of studies have convincingly demonstrated that the extent of tumour infiltration by cytotoxic T cells is a critical factor determining the natural progression of diverse types of cancers[1–4,9–11]. A landmark study showed that the type, density and location of cytotoxic T cells within tumours enabled better prediction of patient survival than histopathological methods used for staging of cancers[1]. Strong infiltration of both the tumour centre and the invasive tumour margin by cytotoxic T cells (which express the CD8 surface marker) was shown to correlate with a favourable prognosis, regardless of the local extent of tumour invasion and spread to local lymph nodes. Conversely, weak *in situ* expansion of CD8 T cells correlated with a poor prognosis even in patients with minimal tumour invasion[1]. However, in the majority of patients this natural defence mechanism is severely blunted by immunosuppressive cell populations recruited to the tumour microenvironment, including regulatory T cells, immature myeloid cell populations and tumour-associated macrophages[4,12–14]. Highly complex interactions among a variety of different cell types in the tumour microenvironment—including tumour cells, immune cells and stromal cells—therefore contribute to clinical outcome.

The critical role of T cells in immune-mediated control of cancers is further underscored by therapeutic benefit following administration of monoclonal antibodies targeting inhibitory receptors on T cells, CTLA-4 and PD-1[15–18]. Clinical benefit is enhanced by co-administration of antibodies targeting CTLA-4 and PD-1[19,20]. Particularly notable is the finding that such antibodies can induce durable responses in a subset of patients with advanced disease. However, many of the regulatory pathways in T cells that result in loss of function within immunosuppressive tumour microenvironments remain unknown.

Immune cells perform complex surveillance functions throughout the body and interact with many different types of cells in distinct tissue microenvironments. Therapeutic targets for modulating immune responses are typically identified *in vitro* and tested in animal models at a late stage of the process. We postulated that the complex interactions of immune cells within tissues, many of which do not occur *in vitro*, offer untapped opportunities for therapeutic intervention. Here we have addressed the challenge of how targets for immune modulation can be systematically discovered *in vivo*.

Design of *in vivo* discovery approach

Pooled shRNA libraries have been shown to be powerful discovery tools[21–23]. We reasoned that shRNAs capable of restoring CD8 T-cell function can be systematically discovered *in vivo* by taking advantage of the extensive proliferative capacity of T cells following triggering of the TCR by a tumour-associated antigen. When introduced into T cells, only a small subset of shRNAs from a pool will restore T-cell proliferation, resulting in their enrichment within tumours. Over-representation of active shRNAs within a pool can be quantified by deep sequencing of the shRNA cassette from tumours and secondary lymphoid organs (Fig. 1a).

We chose to study B16 melanoma, an aggressive tumour that is difficult to treat[24]. Melanoma cells expressed the surrogate tumour antigen ovalbumin (Ova), which is recognized by CD8 T cells from OT-I T-cell receptor transgenic mice[25,26]. Initial experiments showed that such a screen could also be performed with pmel-1 T cells that recognize gp100, an endogenous melanoma antigen[27], but the signal/noise ratio was lower for pmel-1 T cells owing to smaller T-cell populations in tumours. Naive T cells are difficult to infect with lentiviral vectors, and we therefore pretreated T cells for two days with the homeostatic cytokines IL-7 and IL-15 before spin infection with shRNA pools in a lentiviral vector. Successful transduction was monitored by surface expression of the Thy1.1 reporter (Extended Data Fig. 1a). T cells were injected into B6 mice bearing day 14 B16-Ova tumours. Seven days later, T cells were purified from tumours and secondary lymphoid

[1]Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. [2]David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [3]Whitehead Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [5]Novartis Institutes for Biomedical Research, Cambridge, Massachusetts 02139, USA. [6]Genomics Institute of the Novartis Research Foundation, San Diego, California 92121, USA. †Present address: Jounce Therapeutics, Cambridge, Massachusetts 02138, USA.
*These authors contributed equally to this work.

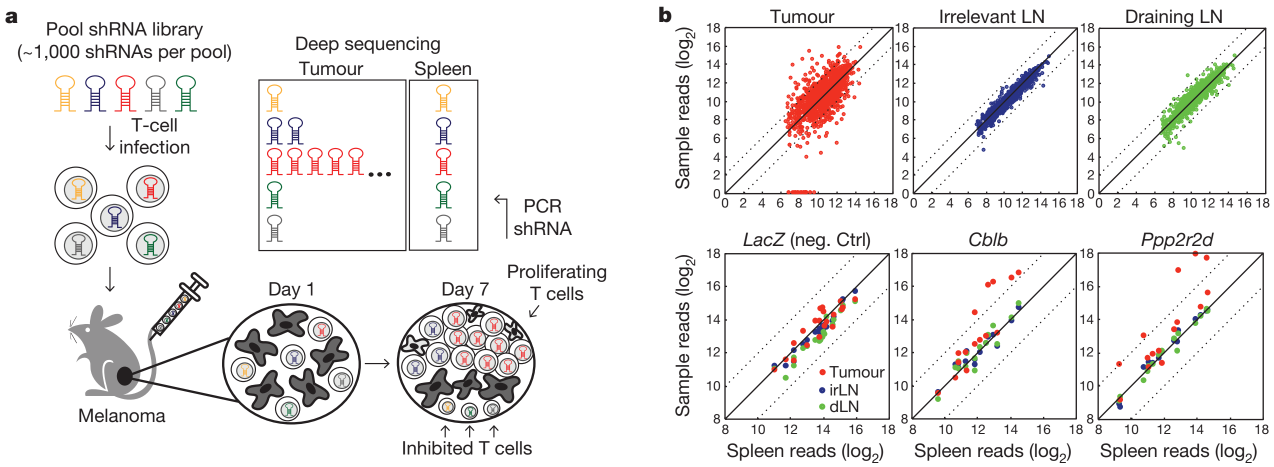


Figure 1 | *In vivo* RNAi discovery of immunotherapy targets. **a**, *In vivo* discovery approach for negative regulators of T-cell function in tumours. T cells infected with shRNA libraries were injected into tumour-bearing mice; shRNAs that enabled T-cell accumulation in tumours were identified by deep sequencing of the shRNA cassette from purified T cells. **b**, Deep sequencing data: shRNA sequence reads from tumours, irrelevant (irLN) and draining lymph nodes (dLN) versus spleen. Upper row, sequence reads for all genes in a pool; lower row, individual genes (*LacZ*, negative control). Dashed lines indicate a deviation by $\log_2$ from diagonal.

organs (spleen, tumour-draining and irrelevant lymph nodes) for isolation of genomic DNA, followed by PCR amplification of the shRNA cassette (Extended Data Fig. 1b). The representation of shRNAs was then quantified in different tissues by Illumina sequencing.

In vivo shRNA pool screens

Two large screens were performed, with the first focusing on genes overexpressed in dysfunctional T cells (T-cell anergy or exhaustion; 255 genes, 1,275 shRNAs divided into two pools), and the second on kinases/phosphatases (1,307 genes, 6,535 shRNAs divided into seven pools) (Table 1a). In these primary screens, each gene was represented by approximately five shRNAs (it is common that only one or two of such shRNAs have sufficient activity in pooled screens). We observed multiple distinct *in vivo* phenotypes. For certain genes, shRNAs were over-represented in all tested tissues compared to the starting T-cell population (for example, SHP-1), indicative of enhanced proliferation independent of TCR recognition of a tumour antigen. For other genes, there was a selective loss of shRNAs within tumours (for example, ZAP-70, a critical kinase in the T-cell activation pathway). We focused our analysis on genes whose shRNAs showed substantial over-representation in tumour but not spleen, a secondary lymphoid organ. Substantial T-cell accumulation in tumours was observed for a number of shRNAs, despite the immunosuppressive environment. For secondary screens, we created focused pools in which each candidate gene was represented by approximately 15 shRNAs. Primary data from this analysis are shown for three genes in Fig. 1b: *LacZ* (negative control), *Cblb* (an E3 ubiquitin ligase that induces T-cell receptor internalization)[28] and *Ppp2r2d* (not previously studied in T cells). For both *Ppp2r2d* and *Cblb*, five shRNAs were substantially increased in tumours (red) compared to spleen, whereas no enrichment was observed for *LacZ* shRNAs. Overall, 43 genes met the following criteria: ≥ fourfold enrichment for three or more shRNAs in tumours compared to spleen (Table 1a and Extended Data Fig. 1c, d). The set included gene products previously identified as inhibitors of T-cell receptor signalling (including *Cblb*, *Dgka*, *Dgkz*, *Ptpn2*), as well as other well-known inhibitors of T-cell function (for example, *Smad2*, *Socs1*, *Socs3*, *Egr2*), validating our approach (Table 1b and Extended Data Table 1)[29–31].

Target validation

We next confirmed at a cellular level that these shRNAs induce T-cell accumulation in tumours. OT-I T cells were infected with lentiviral vectors driving expression of a single shRNA and a reporter protein (Thy1.1 or one of four different fluorescent proteins), and after seven days the frequency of shRNA-transduced T cells was quantified in tumours, spleens and lymph nodes by flow cytometry. When the control *LacZ* shRNA was expressed in CD8 T cells, the frequency of shRNA-expressing CD8 T cells was lower in tumours compared to spleen (~twofold). In contrast, experimental shRNAs induced accumulation

Table 1 | Summary of primary and secondary shRNA screens

a		T-cell dysfunction	Kinase/phosphatase	shRNA enrichment in tumour
First screen	Genes	255	1,307	4–10-fold: 123
	shRNAs	1,275	6,535	10–20-fold: 17
	Candidate genes	32	82	>20-fold: 1
Second screen	Genes	32	43	4–10-fold: 191
	shRNAs	480	645	10–20-fold: 27
	Candidate genes	17	26	>20-fold: 1

b	Function	Genes
	Inhibition of TCR signalling	*Cblb*, *Dgka*, *Dgkz*, *Fyn*, *Inpp5b*, *Ppp3cc*, *Ptpn2*, *Stk17b*, *Tnk1*
	Phosphoinositol metabolism	*Dgka*, *Dgkz*, *Impk*, *Inpp5b*, *Sbf1*
	Inhibitory cytokine signalling pathways	*Smad2*, *Socs1*, *Socs3*
	AMP signalling, Inhibition of mTOR	*Entpd1*, *Prkab2*, *Nuak*
	Cell cycle	*Cdkn2a*, *Pkd1*, *Ppp2r2d*
	Actin and microtubules	*Arhgap5*, *Mast2*, *Rock1*
	Potential nuclear functions	*Blvrb*, *Egr2*, *Impk*, *Jun*, *Ppm1g*
	Role in cancer cells	*Alk*, *Arhgap5*, *Eif2ak3*, *Hipk1*, *Met*, *Nuak*, *Pdzk1ip1*, *Rock1*, *Yes1*

a, T-cell dysfunction and kinase/phosphatase screens. Listed are numbers of genes, shRNAs in each gene set and identified candidate genes. Genes were considered positive in secondary screens when ≥ 3 shRNAs showed ≥ fourfold enrichment in tumour relative to spleen. **b**, Functional classification of candidate genes from secondary screens.

of CD8 T cells in tumours but not in the spleen (Fig. 2a and Extended Data Fig. 2a). T-cell accumulation in tumours was more than tenfold relative to spleen for seven of these genes. The strongest phenotype was observed with shRNAs targeting *Ppp2r2d*, a regulatory subunit of the family of PP2A phosphatases[32]. A *Ppp2r2d* shRNA not only induced accumulation of OT-I CD8 T cells, but also CD4 T cells (from TRP-1 TCR transgenic mice)[33], with T-cell numbers in tumours being significantly higher when *Ppp2r2d* rather than *LacZ* shRNA was expressed (36.3-fold for CD8; 16.2-fold for CD4 T cells) (Fig. 2b). CD8 T-cell accumulation correlated with the degree of *Ppp2r2d* knockdown, and two *Ppp2r2d* shRNAs with the highest *in vivo* activity induced the lowest levels of *Ppp2r2d* messenger RNA (Extended Data Fig. 2b). *Ppp2r2d* knockdown was also confirmed at the protein level using a quantitative mass spectrometry approach (Fig. 2e). *Ppp2r2d* shRNA activity was specific because the phenotype was reversed when a *Ppp2r2d* complementary DNA (with wild-type protein sequence, but mutated DNA sequence at the shRNA binding site) was co-introduced with the *Ppp2r2d* shRNA (Fig. 2c and Extended Data Fig. 3). Furthermore, OT-I CD8 T cells overexpressed *Ppp2r2d* in tumours compared to spleen (in the absence of any shRNA expression), indicating that it is an intrinsic component of the signalling network inhibiting T-cell function in tumours (Fig. 2d). Microarray analysis of tumour-infiltrating T cells expressing different shRNAs showed that each shRNA induced a largely distinct set of gene expression changes, indicating that improved T-cell function in tumours can be mediated through a number of different intracellular pathways (Extended Data Fig. 4).

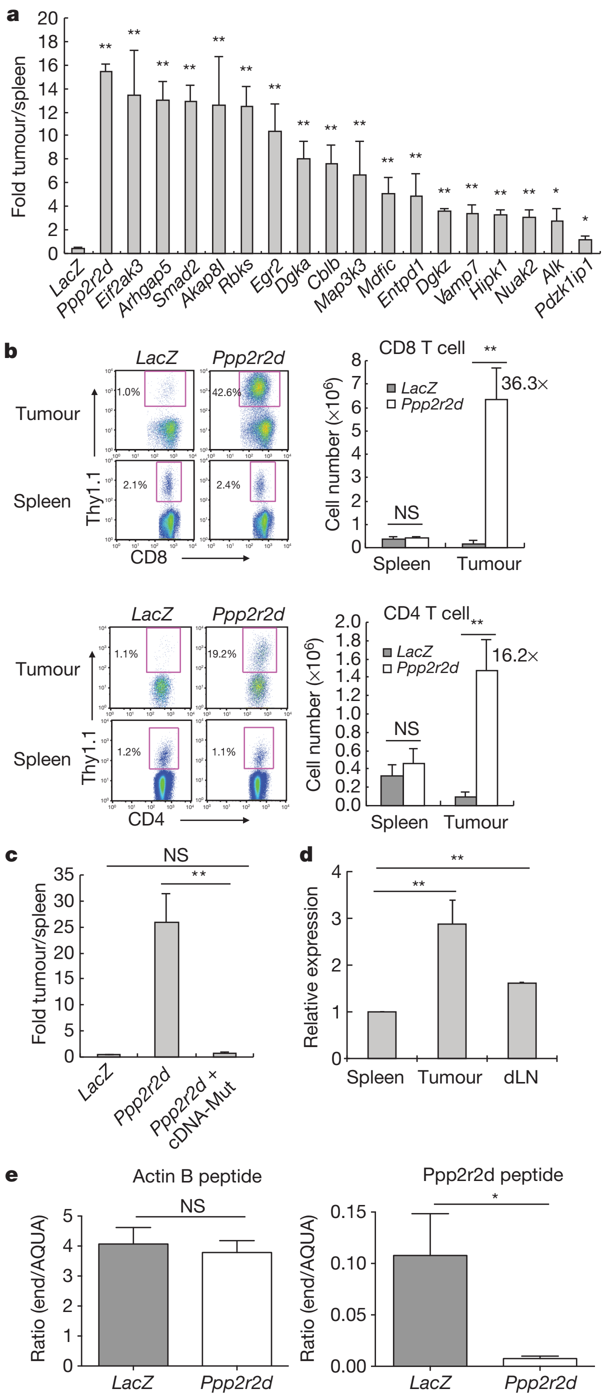


Figure 2 | shRNA-driven accumulation of T cells in B16 melanoma. **a**, CD8 (OT-I) T-cell enrichment in tumours relative to spleen ($n = 3$). **b**, Enrichment of *Ppp2r2d*-silenced CD8 (OT-I) or CD4 (TRP1) T cells (Thy1.1$^+$ cells) in tumour versus spleen. **c**, Reversal of shRNA-induced phenotype by *Ppp2r2d* cDNA with mutated shRNA binding site. NS, not significant. **d**, Quantitative PCR for *Ppp2r2d* mRNA in tumour-infiltrating OT-I T cells (day 7). **e**, Ppp2r2d protein quantification by mass spectrometry with labelled synthetic peptides (AQUA, ratio of endogenous to AQUA peptides). Representative data from two independent experiments (**a**–**d**); Two-sided student's *t*-test, $^{*}P \leq 0.05$, $^{**}P \leq 0.01$; mean ± s.d.

Cellular mechanisms for *Ppp2r2d*

We next examined the cellular mechanisms driving T-cell accumulation by a *Ppp2r2d* shRNA in tumours, specifically T-cell infiltration, proliferation and apoptosis. T-cell infiltration into tumours was assessed by transfer of OT-I CD8 T cells labelled with a cytosolic dye (carboxyfluorescein succinimidyl ester, CFSE). No differences were observed in the frequency of *Ppp2r2d* or *LacZ* shRNA-transduced CD8 T cells in tumours on day 1, indicating no substantial effect on T-cell infiltration (Fig. 3a). However, analysis of later time points (days 3–7) demonstrated a higher degree of proliferation (based on CFSE dilution) by *Ppp2r2d* compared to *LacZ* shRNA-transduced T cells (Fig. 3b and Extended Data Fig. 5a). The action of *Ppp2r2d* was downstream of T-cell receptor activation because T-cell proliferation was enhanced in tumours and to a lesser extent in tumour-draining lymph nodes (Extended Data Fig. 5a). In contrast, no proliferation was observed in irrelevant lymph nodes or the spleen where the relevant antigen was not presented to T cells (data not shown). Substantial T-cell proliferation was even observed for *LacZ* shRNA-transduced T cells (complete dilution of CFSE dye by day 7), despite the presence of small numbers of such cells in tumours. This indicated that *LacZ* shRNA-transduced T cells were lost by apoptosis. Indeed, a larger percentage of tumour-infiltrating T cells were labelled with an antibody specific for active caspase 3 when the *LacZ* control shRNA (rather than *Ppp2r2d* shRNA) was expressed (Fig. 3c and Extended Data Fig. 5b). Furthermore, co-culture of CD8 T cells with B16-Ova tumour cells showed that the majority of *LacZ* shRNA-expressing T cells became apoptotic (65.7%), whereas most *Ppp2r2d* shRNA-transduced T cells were viable (89.5%, Fig. 3d).

These results indicated the possibility that *Ppp2r2d* shRNA-transduced CD8 T cells may be able to proliferate and survive even when they recognize their antigen directly presented by B16-Ova tumour cells. This idea was tested by implantation of tumour cells into *B2m*$^{-/-}$ mice which are deficient in expression of MHC class I proteins[34]. In such mice, only tumour cells of the host, but not professional antigen-presenting cells, could present tumour antigens to T cells. Indeed, *Ppp2r2d* shRNA-transduced OT-I CD8 T cells showed massive accumulation within B16-Ova tumours in *B2m*$^{-/-}$ mice (Fig. 3e) whereas very small numbers of T cells were present in contralateral B16 tumours that lacked expression of the Ova antigen. *Ppp2r2d*-silenced T cells could therefore effectively proliferate and survive in response to tumour cells, despite a lack of suitable co-stimulatory signals and an inhibitory microenvironment.

Ex vivo analysis of tumour-infiltrating T cells at a single-cell level using a nanowell device[35,36] also demonstrated that *Ppp2r2d* silencing increased cytokine production by T cells (Fig. 4a–c). T cells were activated

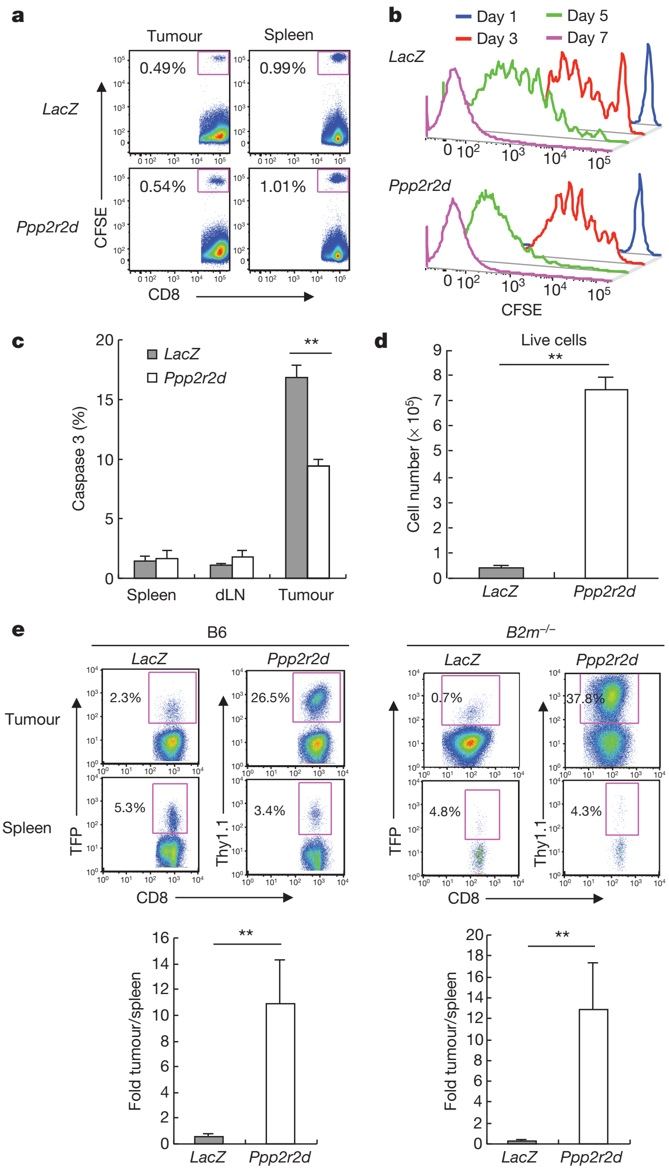


Figure 3 | Changes in T-cell function induced by *Ppp2r2d* shRNA. **a**, Tumour infiltration at 24 h by CFSE-labelled OT-I T cells. **b**, Enhanced proliferation by *Ppp2r2d*-silenced T cells (CFSE dilution). **c**, **d**, Reduced apoptosis by *Ppp2r2d*-silenced OT-I T cell in tumours (**c**, activated caspase-3) or during 3-day co-culture with B16-Ova tumour cells (**d**, annexin V). **e**, *Ppp2r2d*-silencing induced T-cell expansion even when MHC class I expression was restricted to tumour cells; T-cell transfer into C57BL/6 or *B2m*$^{-/-}$ mice with B16-Ova tumours. Data representative of two independent trials ($n = 3$; $^{**}P \leq 0.01$, two-sided student's t-test); mean ± s.d.

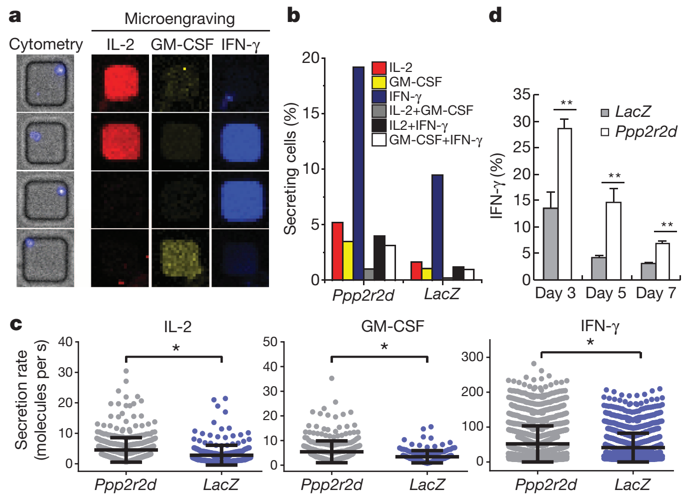


Figure 4 | Cytokine secretion by gene-silenced tumour-infiltrating T cells. **a–c**, *Ex vivo* analysis of cytokine production by tumour-infiltrating OT-I T cells at a single-cell level using a nanowell device (84,672 wells of picolitre volume). **a**, Representative single cells in nanowells and corresponding patterns of cytokine secretion. **b**, Percentage of T cells secreting indicated cytokines. **c**, Cytokine secretion rates calculated from standard curves (mean ± s.d., $^{*}P < 0.05$, Mann–Whitney U-test). **d**, Intracellular IFN-γ staining for tumour-infiltrating *Ppp2r2d*-silenced T cells, representative of two independent experiments ($n = 3$, $^{**}P \leq 0.01$, two-sided student's t-test); mean ± s.d.

for 3 h by CD3/CD28 antibodies on lipid bilayers, followed by 1 h cytokine capture on antibody-coated slides. CD8 T cells showed a higher secretion rate for interferon-γ, interleukin-2 and granulocyte–macrophage colony-stimulating factor (IFN-γ, IL-2 and GM-CSF, respectively) and a larger fraction of T cells secreted more than one cytokine (Fig. 4b, c). The presence of larger numbers of IFN-γ-producing T cells was confirmed by intracellular cytokine staining (Fig. 4d and Extended Data Fig. 5c).

PP2A represents a family of phosphatase complexes composed of catalytic, scaffolding and regulatory subunits. Cellular localization and substrate specificity are determined by one of many regulatory subunits, of which Ppp2r2d is a member[32]. Ppp2r2d directs PP2A to Cdk1 substrates during interphase and anaphase; it thereby inhibits entry into mitosis and induces exit from mitosis[37]. PP2A also has a gatekeeper role for BAD-mediated apoptosis. Phosphorylated BAD is sequestered in its inactive form in the cytosol by 14-3-3, whereas dephosphorylated BAD is targeted to mitochondria where it causes cell death by binding Bcl-X_L and Bcl-2[38]. PP2A phosphatases have also been shown to interact with the cytoplasmic domains of CD28 and CTLA-4 as well as Carma1 (upstream of the NF-κB pathway)[39,40], but it is not known which regulatory subunits are required for these activities. Anti-Ppp2r2d antibodies suitable for the required biochemical studies are not currently available.

Enhanced anti-tumour immunity

Finally, we assessed the ability of a *Ppp2r2d* shRNA to enhance the efficacy of adoptive T-cell therapy. B16-Ova tumour cells (2×10^5) were injected subcutaneously into B6 mice. On day 12, mice bearing tumours of similar size were divided into seven groups, either receiving no T cells, 2×10^6 shRNA-transduced TRP-1 CD4 T cells, 2×10^6 shRNA-infected OT-I CD8 T cells, or both CD4 and CD8 T cells (days 12 and 17). The modest anti-tumour activity of OT-I CD8 T cells (expressing the control *LacZ* shRNA) is consistent with published data[41]. *Ppp2r2d*-silencing improved the therapeutic activity of both CD4 and CD8 T cells (Fig. 5a, b). A *Ppp2r2d* shRNA also enhanced anti-tumour responses when introduced into T cells specific for the endogenous melanoma antigens gp100 (pmel-1 CD8 T cells) and TRP-1 (TRP-1 CD4 T cells) (Fig. 5c). gp100 is a relevant antigen in human melanoma, and a clinical trial in which a gp100-specific TCR (isolated from HLA-A2 transgenic mice) was introduced into peripheral blood T cells demonstrated therapeutic benefit in a subset of patients[42].

Ppp2r2d-silenced T cells acquired an effector phenotype in tumours (Extended Data Fig. 6a) and >30% of the cells expressed granzyme B (Extended Data Fig. 7a). Consistent with greatly increased numbers of such effector T cells in tumours (Extended Data Fig. 7b), terminal deoxynucleotidyl transferase dUTP nick end labelling (TUNEL) demonstrated increased apoptosis in tumours when *Ppp2r2d* rather than *LacZ* shRNA-expressing T cells were present (Extended Data Fig. 7c). B16 melanomas are highly aggressive tumours in part because MHC class I expression is very low. Interestingly, *Ppp2r2d* but not *LacZ* shRNA-expressing T cells significantly increased MHC class I expression (H-2K^b) by tumour cells (Extended Data Fig. 7d), possibly due to the observed increase in IFN-γ secretion by T cells (Fig. 4b-d). A

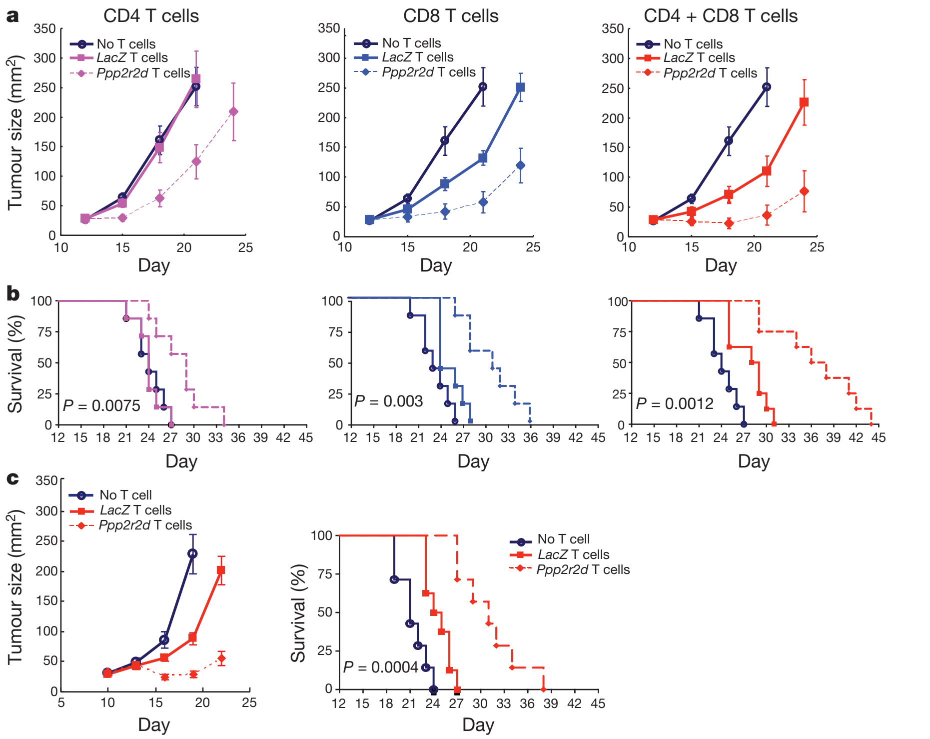


Figure 5 | *Ppp2r2d* silencing enhances anti-tumour activity of CD4 and CD8 T cells. T cells were activated with CD3/CD28 beads and infected with shRNA vectors. **a**, **b**, $CD4^+$ TRP-1 and/or $CD8^+$ OT-I T cells (2×10^6) were transferred (day 12 and 17) into mice bearing day 12 B16-Ova tumours. Tumour burden (**a**) and survival (**b**) were assessed. **c**, $CD4^+$ TRP-1 and $CD8^+$ pmel-1 T cells (3×10^6 each) were transferred (day 10 and 15) into mice with day 10 B16 tumours. Representative of two independent experiments ($n = 7$–9 mice per group), survival analysed using log-rank (Mantel–Cox) test; mean ± s.e.m.

Ppp2r2d shRNA did not reduce expression of inhibitory PD-1 or LAG-3 receptors on tumour-infiltrating T cells, demonstrating that its mechanism of action is distinct from these known negative regulators of T-cell function (Extended Data Fig. 6b). This finding suggests combination approaches targeting these intracellular and cell surface molecules.

Discussion

These results establish the feasibility of *in vivo* discovery of novel targets for immunotherapy in complex tissue microenvironments. We show that it is possible to discover genes with differential action across tissues, as exemplified by T-cell accumulation in tumours compared to secondary lymphoid organs. For genes with tissue-selective action, T-cell proliferation and survival are likely to be under the control of the T-cell receptor and therefore do not occur in tissues lacking presentation of a relevant antigen. Many variations of the approach presented here can be envisioned to investigate control of particular immune cell functions *in vivo*. For example, fluorescent reporters for expression of cytokines or cytotoxic molecules (granzyme B, perforin) could be integrated into our approach to discover genes that control critical T-cell effector functions in tumours.

Targeting of key regulatory switches may offer new approaches to modify the activity of T cells in cancer and other pathologies. For example, recent clinical trials have shown that transfer of genetically modified T cells can result in substantial anti-tumour activity[43–46]. The efficacy of such T-cell-based therapies could be enhanced by shRNA-mediated silencing of genes that inhibit T-cell function in the tumour microenvironment.

METHODS SUMMARY

***In vivo* shRNA screening.** Nine shRNA pools (approximately 5 shRNAs per gene) were created and subcloned into the pLKO-Thy1.1 lentiviral vector. Each pool also included 85 negative-control shRNAs. OT-I T cells were cultured with IL-7 ($5\,\mathrm{ng\,ml^{-1}}$) and IL-15 ($100\,\mathrm{ng\,ml^{-1}}$); on day 2 cells were spin-infected with lentiviral pools supplemented with protamine sulphate ($5\,\mu\mathrm{g\,ml^{-1}}$) in RetroNectin-coated 24-well plates ($5\,\mu\mathrm{g\,ml^{-1}}$) at a multiplicity of infection (MOI) of 15. Following infection, OT-1 cells were cultured with IL-7 ($2.5\,\mathrm{ng\,ml^{-1}}$), IL-15 ($50\,\mathrm{ng\,ml^{-1}}$) and IL-2 ($2\,\mathrm{ng\,ml^{-1}}$). On day 5, shRNA-transduced T cells were enriched by positive selection using the Thy1.1 surface reporter (StemCell Technologies). T cells (5×10^6) were injected intravenously into C57BL/6 mice bearing day 14 B16-Ova tumours (15 mice per shRNA pool). Seven days later, shRNA-expressing T cells ($CD8^+V\alpha2^+V\beta5^+Thy1.1^+$) were isolated by flow cytometry from tumours, spleens, tumour-draining lymph nodes and irrelevant lymph nodes. Genomic DNA was purified (Qiagen) and deep-sequencing templates were generated by PCR amplification of the shRNA cassette. Representation of shRNAs in each pool was analysed by deep sequencing using an Illumina Genome Analyzer[47].

Secondary screens were performed using focused pools containing approximately 15 shRNAs per gene as well as 85 negative controls. Cut-off in the secondary screen was defined as ≥ 3 shRNAs with ≥ fourfold enrichment in tumour relative to spleen. Screening results were validated at a cellular level by introducing individual shRNAs into T cells, along with a reporter protein (green, teal, red or ametrine fluorescent proteins, Thy1.1). This approach enabled simultaneous testing of five shRNAs in an animal (three mice per group). Proliferation of shRNA-transduced T cells was visualized on the basis of CFSE dilution after 24 h as well as 3, 5 and 7 days.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 April; accepted 31 December 2013.

Published online 29 January 2014.

1. Galon, J. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313,** 1960–1964 (2006).
2. Hamanishi, J. *et al.* Programmed cell death 1 ligand 1 and tumor-infiltrating $CD8^+$ T lymphocytes are prognostic factors of human ovarian cancer. *Proc. Natl Acad. Sci. USA* **104,** 3360–3365 (2007).
3. Mahmoud, S. M. *et al.* Tumor-infiltrating $CD8^+$ lymphocytes predict clinical outcome in breast cancer. *J. Clin. Oncol.* **29,** 1949–1955 (2011).
4. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39,** 782–795 (2013).
5. Matsushita, H. *et al.* Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* **482,** 400–404 (2012).
6. Oble, D. A., Loewe, R., Yu, P. & Mihm, M. C. Jr. Focus on TILs: prognostic significance of tumor infiltrating lymphocytes in human melanoma. *Cancer Immun.* **9,** 3 (2009).
7. DuPage, M., Mazumdar, C., Schmidt, L. M., Cheung, A. F. & Jacks, T. Expression of tumour-specific antigens underlies cancer immunoediting. *Nature* **482,** 405–409 (2012).
8. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* **331,** 1565–1570 (2011).

9. Pagès, F. *et al.* In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer. *J. Clin. Oncol.* **27,** 5944–5951 (2009).
10. Rusakiewicz, S. *et al.* Immune infiltrates are prognostic factors in localized gastrointestinal stromal tumors. *Cancer Res.* **73,** 3499–3510 (2013).
11. Stumpf, M. *et al.* Intraepithelial CD8-positive T lymphocytes predict survival for patients with serous stage III ovarian carcinomas: relevance of clonal selection of T lymphocytes. *Br. J. Cancer* **101,** 1513–1521 (2009).
12. Gabrilovich, D. I. & Nagaraj, S. Myeloid-derived suppressor cells as regulators of the immune system. *Nature Rev. Immunol.* **9,** 162–174 (2009).
13. Shiao, S. L., Ganesan, A. P., Rugo, H. S. & Coussens, L. M. Immune microenvironments in solid tumors: new targets for therapy. *Genes Dev.* **25,** 2559–2572 (2011).
14. Tanchot, C. *et al.* Tumor-infiltrating regulatory T cells: phenotype, role, mechanism of expansion in situ and clinical significance. *Cancer Microenviron.* **6,** 147–157 (2013).
15. Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* **363,** 711–723 (2010).
16. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* **366,** 2443–2454 (2012).
17. Brahmer, J. R. *et al.* Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N. Engl. J. Med.* **366,** 2455–2465 (2012).
18. Leach, D. R., Krummel, M. F. & Allison, J. P. Enhancement of antitumor immunity by CTLA-4 blockade. *Science* **271,** 1734–1736 (1996).
19. Wolchok, J. D. *et al.* Nivolumab plus ipilimumab in advanced melanoma. *N. Engl. J. Med.* **369,** 122–133 (2013).
20. Curran, M. A., Montalvo, W., Yagita, H. & Allison, J. P. PD-1 and CTLA-4 combination blockade expands infiltrating T cells and reduces regulatory T and myeloid cells within B16 melanoma tumors. *Proc. Natl Acad. Sci. USA* **107,** 4275–4280 (2010).
21. Westbrook, T. F. *et al.* A genetic screen for candidate tumor suppressors identifies REST. *Cell* **121,** 837–848 (2005).
22. Zender, L. *et al.* An oncogenomics-based *in vivo* RNAi screen identifies tumor suppressors in liver cancer. *Cell* **135,** 852–864 (2008).
23. Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *Proc. Natl Acad. Sci. USA* **105,** 20380–20385 (2008).
24. Fidler, I. J. Biological behavior of malignant melanoma cells correlated to their survival *in vivo*. *Cancer Res.* **35,** 218–224 (1975).
25. Hogquist, K. A. *et al.* T cell receptor antagonist peptides induce positive selection. *Cell* **76,** 17–27 (1994).
26. Bellone, M. *et al.* Relevance of the tumor antigen in the validation of three vaccination strategies for melanoma. *J. Immunol.* **165,** 2651–2656 (2000).
27. Overwijk, W. W. *et al.* Tumor regression and autoimmunity after reversal of a functionally tolerant state of self-reactive $CD8^+$ T cells. *J. Exp. Med.* **198,** 569–580 (2003).
28. Paolino, M. & Penninger, J. M. Cbl-b in T-cell activation. *Semin. Immunopathol.* **32,** 137–148 (2010).
29. Zheng, Y., Zha, Y. & Gajewski, T. F. Molecular regulation of T-cell anergy. *EMBO Rep.* **9,** 50–55 (2008).
30. Doody, K. M., Bourdeau, A. & Tremblay, M. L. T-cell protein tyrosine phosphatase is a key regulator in immune cell signaling: lessons from the knockout mouse model and implications in human disease. *Immunol. Rev.* **228,** 325–341 (2009).
31. Tamiya, T., Kashiwagi, I., Takahashi, R., Yasukawa, H. & Yoshimura, A. Suppressors of cytokine signaling (SOCS) proteins and JAK/STAT pathways: regulation of T-cell inflammation by SOCS1 and SOCS3. *Arterioscler. Thromb. Vasc. Biol.* **31,** 980–985 (2011).
32. Barr, F. A., Elliott, P. R. & Gruneberg, U. Protein phosphatases and the regulation of mitosis. *J. Cell Sci.* **124,** 2323–2334 (2011).
33. Muranski, P. *et al.* Tumor-specific Th17-polarized cells eradicate large established melanoma. *Blood* **112,** 362–373 (2008).
34. Koller, B. H., Marrack, P., Kappler, J. W. & Smithies, O. Normal development of mice deficient in beta 2M, MHC class I proteins, and $CD8^+$ T cells. *Science* **248,** 1227–1230 (1990).
35. Torres, A. J., Contento, R. L., Gordo, S., Wucherpfennig, K. W. & Love, J. C. Functional single-cell analysis of T-cell activation by supported lipid bilayer-tethered ligands on arrays of nanowells. *Lab Chip* **13,** 90–99 (2013).
36. Han, Q., Bradshaw, E. M., Nilsson, B., Hafler, D. A. & Love, J. C. Multidimensional analysis of the frequencies and rates of cytokine secretion from single cells by quantitative microengraving. *Lab Chip* **10,** 1391–1400 (2010).
37. Mochida, S., Maslen, S. L., Skehel, M. & Hunt, T. Greatwall phosphorylates an inhibitor of protein phosphatase 2A that is essential for mitosis. *Science* **330,** 1670–1673 (2010).
38. Chiang, C. W. *et al.* Protein phosphatase 2A dephosphorylation of phosphoserine 112 plays the gatekeeper role for BAD-mediated apoptosis. *Mol. Cell. Biol.* **23,** 6350–6362 (2003).
39. Chuang, E. *et al.* The CD28 and CTLA-4 receptors associate with the serine/threonine phosphatase PP2A. *Immunity* **13,** 313–322 (2000).
40. Eitelhuber, A. C. *et al.* Dephosphorylation of Carma1 by PP2A negatively regulates T-cell activation. *EMBO J.* **30,** 594–605 (2011).
41. Tao, J. *et al.* JNK2 negatively regulates $CD8^+$ T cell effector function and anti-tumor immune response. *Eur. J. Immunol.* **37,** 818–829 (2007).
42. Johnson, L. A. *et al.* Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood* **114,** 535–546 (2009).
43. Brenner, M. K. & Heslop, H. E. Adoptive T cell therapy of cancer. *Curr. Opin. Immunol.* **22,** 251–257 (2010).
44. Turtle, C. J., Hudecek, M., Jensen, M. C. & Riddell, S. R. Engineered T cells for anti-cancer therapy. *Curr. Opin. Immunol.* **24,** 633–639 (2012).
45. Kalos, M. & June, C. H. Adoptive T cell transfer for cancer immunotherapy in the era of synthetic biology. *Immunity* **39,** 49–60 (2013).
46. Restifo, N. P., Dudley, M. E. & Rosenberg, S. A. Adoptive immunotherapy for cancer: harnessing the T cell response. *Nature Rev. Immunol.* **12,** 269–281 (2012).
47. Ashton, J. M. *et al.* Gene sets identified with oncogene cooperativity analysis regulate *in vivo* growth and survival of leukemia stem cells. *Cell Stem Cell* **11,** 359–372 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the National Institutes of Health (Transformative Research Award 1R01CA173750 to K.W.W.), the Melanoma Research Alliance (to K.W.W.), the DF/HCC–MIT Bridge Project and the Lustgarten Foundation (to K.W.W., J.C.L. and H.L.P.), Novartis Institutes of Biomedical Research (to K.W.W.), the Koch Institute Support Grant P30-CA14051 from the National Cancer Institute, the American Cancer Society John W. Thatcher, Jr Postdoctoral Fellowship in Melanoma Research (to D.R.S.), the Terri Brodeur Breast Cancer Foundation Postdoctoral Fellowship (to P.Z.) and a NIH T32 grant (AI07386 to D.A.A.A.).

Author Contributions K.W.W., P.Z., S.J.T., G.D. and H.C. contributed to the overall study design; K.W.W., P.Z. and D.R.S. designed experiments; P.Z., D.A.A.A. and H.C. developed procedure for lentiviral infection of T cells and optimized approaches for adoptive T-cell therapy; P.Z., D.R.S. and D.A.A.A. performed shRNA screen; G.S.C., D.E.R. and N.H. provided pooled shRNA library and advice on shRNA screen; Y.N. and G.D. provided B16-Ova cell line and advice on tumour model; A.J.T. and J.C.L. performed nano-well analysis of cytokine production, V.C. and S.J.T. performed histological studies, W.P. performed protein quantification by mass spectrometry, S.K.D. and H.L.P. provided mouse models; J.B., K.E. and J.L. performed microarray analysis; K.W.W., P.Z. and D.R.S. wrote the paper.

Author Information The access number for microarray data is GSE53388 in the Genomic Spatial Event (GSE) database. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.W.W. (kai_wucherpfennig@dfci.harvard.edu).

OPEN
doi:10.1038/nature12959

An environmental bacterial taxon with a large and distinct metabolic repertoire

Micheal C. Wilson[1,2]*, Tetsushi Mori[3]*, Christian Rückert[4], Agustinus R. Uria[1,2], Maximilian J. Helf[1,2], Kentaro Takada[5], Christine Gernert[6], Ursula A. E. Steffens[2], Nina Heycke[2], Susanne Schmitt[7], Christian Rinke[8], Eric J. N. Helfrich[1,2], Alexander O. Brachmann[1], Cristian Gurgui[2], Toshiyuki Wakimoto[9], Matthias Kracht[2], Max Crüsemann[2], Ute Hentschel[6], Ikuro Abe[9], Shigeki Matsunaga[5], Jörn Kalinowski[4], Haruko Takeyama[3] & Jörn Piel[1,2]

Cultivated bacteria such as actinomycetes are a highly useful source of biomedically important natural products. However, such 'talented' producers represent only a minute fraction of the entire, mostly uncultivated, prokaryotic diversity. The uncultured majority is generally perceived as a large, untapped resource of new drug candidates, but so far it is unknown whether taxa containing talented bacteria indeed exist. Here we report the single-cell- and metagenomics-based discovery of such producers. Two phylotypes of the candidate genus '*Entotheonella*' with genomes of greater than 9 megabases and multiple, distinct biosynthetic gene clusters co-inhabit the chemically and microbially rich marine sponge *Theonella swinhoei*. Almost all bioactive polyketides and peptides known from this animal were attributed to a single phylotype. '*Entotheonella*' spp. are widely distributed in sponges and belong to an environmental taxon proposed here as candidate phylum 'Tectomicrobia'. The pronounced bioactivities and chemical uniqueness of '*Entotheonella*' compounds provide significant opportunities for ecological studies and drug discovery.

More than half of the known natural products with antimicrobial, antitumour or antiviral activity are of bacterial origin[1]. Most of these compounds were isolated from cultivated representatives of only five bacterial groups: filamentous actinomycetes, Myxobacteria, Cyanobacteria, and members of the genera *Pseudomonas* and *Bacillus*. Uncultivated bacteria, which are proposed to form 70% of all known prokaryotic phyla[2], represent a particularly promising source for new, chemically prolific taxa. However, except for individual biosynthetic pathways reported from environmental sources[3,4], the true metabolic potential of these microbes remains unexplored. Two such pathways, involved in the production of onnamide- and theopederin-type polyketides[5] and ribosomal peptides of the polytheonamide group[6] (Fig. 1), were previously discovered in the marine sponge *Theonella swinhoei*. Like many other sponges, this animal harbours a massive consortium of uncultivated bacteria belonging to hundreds of distinct phylotypes[7–9]. *T. swinhoei* is the source of exceptionally diverse natural products and forms distinct chemotypes; samples of the sponge collected from different locations have largely non-overlapping metabolite profiles. From the onnamide and polytheonamide chemotype occurring at Hachijo Jima, Japan, here termed *T. swinhoei* Y (Y referring to the yellow interior of the sponge), in total more than 40 bioactive polyketides and modified peptides belonging to seven structural classes were isolated (Fig. 1)[10]. As previous work on onnamides and polytheonamides has produced only metagenomic DNA fragments lacking taxonomically diagnostic features, it was unknown which members of the bacterial community are the producers of these compounds.

Attribution of metabolic genes to '*Entotheonella*'

Single-cell analysis has recently emerged as an efficient strategy to correlate the phylogenetic identity of environmental microorganisms with their functional gene repertoire[11–13]. To pinpoint producers in *T. swinhoei* Y, samples enriched in bacteria of different cell densities were prepared by differential centrifugation after sponge collection. When a fraction of higher density (Fig. 2a) was microscopically examined, we found that it contained a highly enriched population of large filamentous bacteria that fluoresce when excited with ultraviolet light (Fig. 2b). The bacteria were morphologically similar to the symbiont '*Candidatus* Entotheonella palauensis' previously reported from a Palauan *Theonella swinhoei* chemotype and suspected as producer of antifungal peptides[14,15]. Scanning electron micrographs (Fig. 2c) revealed the presence of approximately 2- to 3-μm cells linked to each other. These bacteria, as well as those from the low-density fraction containing mostly unicellular bacteria, were sorted individually into 96-well plates by fluorescence-assisted cell sorting (FACS) (Extended Data Fig. 1a), resulting in filamentous and unicellular plates. Subsequently, multiple displacement amplification (MDA) of single bacterial genomes was performed on each well, resulting in DNA product sizes of approximately 10 kb (Extended Data Fig. 1b).

To detect wells containing DNA from the onnamide or polytheonamide producer, primers specific for *onn* and *poy* genes encoding the respective pathways were used in diagnostic PCRs. For both gene clusters, a large number of positive wells were detected among the filamentous plates (Fig. 2d and Extended Data Fig. 1c). Subsequent PCRs with eubacterial and '*Entotheonella*'-specific 16S ribosomal RNA gene primers showed that about half of the wells contained DNA originating from '*Entotheonella*' phylotypes. Overall, from 48 wells of an analysed filamentous plate, 22 wells were positive for the onnamide, 34 for the polytheonamide, and 27 for the '*Entotheonella*' sp. 16S rRNA gene, as confirmed by sequencing of each amplicon. Sixteen of the positive wells showed amplification for all three of the *onn*, *poy* and '*Entotheonella*' sp.

[1]Institute of Microbiology, Eidgenössische Technische Hochschule Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland. [2]Kekulé Institute of Organic Chemistry and Biochemistry, University of Bonn, Gerhard-Domagk-Strasse 1, 53121 Bonn, Germany. [3]Faculty of Science and Engineering, Waseda University Center for Advanced Biomedical Sciences, 2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8480, Japan. [4]Institute for Genome Research and Systems Biology, Center for Biotechnology, Universität Bielefeld, Universitätstrasse 25, 33594 Bielefeld, Germany. [5]Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan. [6]Department of Botany II, Julius-von-Sachs Institute for Biological Sciences, University of Würzburg, Julius-von-Sachs-Platz 3, 97082 Würzburg, Germany. [7]Department of Earth and Environmental Sciences, Palaeontology and Geobiology, Ludwig Maximilians University Munich, Richard-Wagner-Strasse 10, 80333 Munich, Germany. [8]Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [9]Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.
*These authors contributed equally to this work.

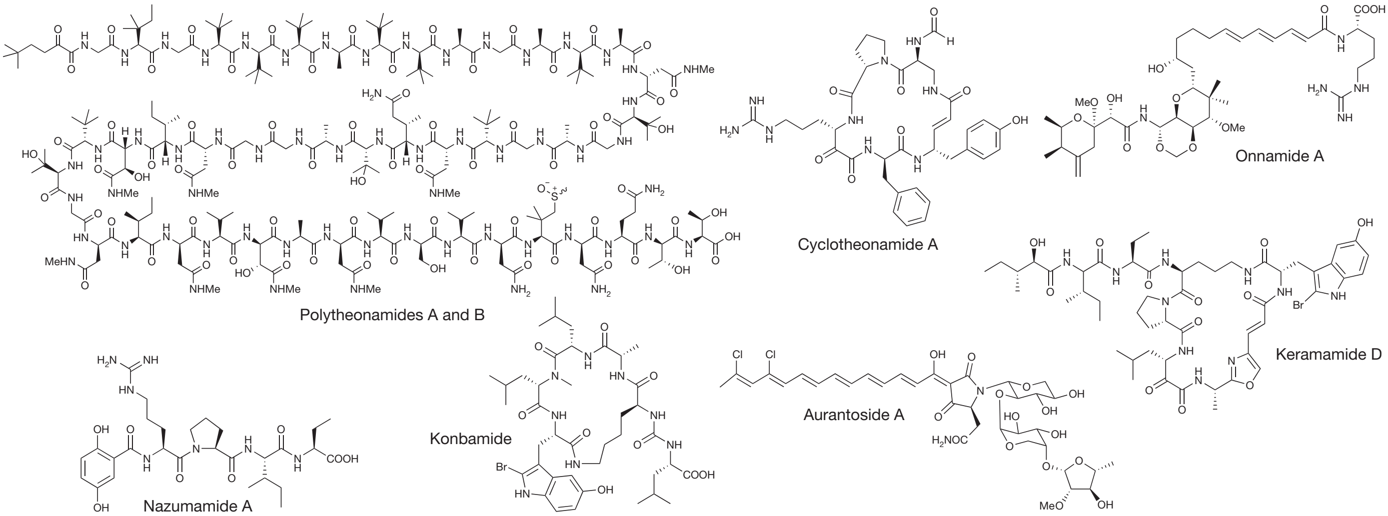


Figure 1 | Representative bioactive natural product families isolated from the sponge *Theonella swinhoei*. Polytheonamides A and B differ in the stereochemistry of the sulphoxide moiety (polytheonamide A shows *S* chirality; polytheonamide B shows *R* chirality).

primer pairs in one or more out of three repetitive PCRs (Fig. 2d). For further analysis, wells positive for all three primer sets were subjected to PCR using eubacterial 16S rRNA gene primers. 16S rRNA genes from '*Entotheonella*' sp. as well as *Escherichia coli* were identified. The *E. coli* amplicon was discarded, as it was also identified in MDA-treated wells that only contained water. Thus, the data suggested '*Entotheonella*' as the source of both the onnamide-type compounds and polytheonamides.

Two chemically distinct '*Entotheonella*' symbionts

As not all wells were positive for all three primer pairs and bacteria might have been overlooked owing to incomplete genome coverage during MDA[16], we wished to validate further our results by metagenomic sequencing. On the filamentous bacterial cell sample, several rounds of Illumina, 454, PacBio, and Sanger sequencing were performed (Supplementary Table 1). Of the sequencing reads, 78.3% assembled to longer contigs, resulting in 18,093 contigs of at least 500 bp. The remaining reads did not show significant overlap, suggesting that the corresponding phylotypes were present only at low concentration. This hypothesis was backed by the observation of a high variance in coverage, ranging between 3.3- and 1,564.7-fold for contigs of at least 2 kb length. Basic Local Alignment Search Tool X (BLASTX) analysis of the contig and scaffold sequences followed by binning based on sequence depth and G + C content revealed two large populations of bacterial DNA with a G + C content around 55% (Supplementary Table 2). A third set of low coverage and low G + C contigs delivered hits against various eukaryotic genomes and was therefore excluded from further analyses (Extended Data Fig. 2a). A more detailed analysis of the filtered data set revealed for most bacterial genes the existence of two highly similar versions (approximately 85–91% nucleotide identity) that resided in virtually syntenous genomic environments encompassing over 4.5 Mb (Extended Data Fig. 3). The overlapping genomic regions included exactly two orthologues of 35 single-copy genes often used as bacterial phylogenic markers (Supplementary Table 3)[17]. These features suggested that the large majority of assembled bacterial sequences belonged to two closely related '*Entotheonella*' variants, termed TSY1 and TSY2, with 97.6% identical 16S rRNA gene sequences and an average G + C content of 55.79% (Extended Data Fig. 2b). The identity of the 16S rRNA genes to that of '*E. palauensis*' was about 97%. Depth analysis also suggested the presence of about 236 kb of DNA belonging to at least one large plasmid (G + C content: 55.11%). Coverage was 60.3-, 24.5- and 278.5-fold for the TSY1 and TSY2 chromosomes and the plasmid(s), respectively (corresponding to a ratio of 1:0.4:5), indicating that TSY1 is the dominant strain (Extended Data Fig. 2b). Both strains possess genomes of similar size that exceed 9 Mb, thus belonging to the largest known prokaryotic genomes (Supplementary Table 2). A remarkably large number of repetitive elements, some present in about 25 to 100 copies, as well as the high degree of similarity of the two genomes prohibited further assembly. To determine completeness of genomes, a core gene group analysis[18] was performed, identifying 62 of 66 core groups for both TSY1 and TSY2. Thus we assume that the protein inventory of both strains was almost completely established.

The search for metabolic genes in this data set revealed complete sets of *onn* and *poy* genes on the plasmid-derived contigs. In addition, an unexpectedly high number of further gene clusters for polyketide and ribosomal or non-ribosomal peptide biosynthesis were identified on the chromosomal sequences. To allow for prediction of the corresponding metabolites, sequence gaps within most of these loci were filled by paired-end sequencing of 3- and 8-kb libraries and by combinatorial or targeted PCR, resulting in at least 28 biosynthetic gene clusters on 31 scaffolds (Extended Data Fig. 4 and Supplementary Table 4). For many non-ribosomal peptide synthetase (NRPS) clusters, bioinformatic predictions based on enzyme colinearity rules[19], substrate recognition motifs[20–22], and the presence of genes for non-proteinogenic amino acid biosynthesis (Supplementary Table 5), revealed known bioactive peptides from Japanese *T. swinhoei* as the best structural hits. Specifically, we identified virtually perfect matches for the cyclotheonamides, keramamides and nazumamide A. In addition, we identified a konbamide A-type[23] cluster in which five of the six NRPS modules are present and colinear with the compound structure, but two ORF insertions disrupt the NRPS architecture, suggesting that the cluster is an inactive evolutionary relic. Consistent with this, members of the onnamide, polytheonamide, keramamide, and cyclotheonamide compound families were detected using high-resolution mass spectroscopy (HRMS) in extracts of our sponge specimens and enriched filamentous cell fractions, but we were unable to detect the konbamides (Supplementary Tables 6 and 7, and Extended Data Fig. 5). Taking together the combined bioinformatic and chemical analyses, candidate gene clusters existed for all known peptide and polyketide families including onnamides and polytheonamides, except for the aurantosides. In addition to these attributable genes, loci for at least 14 peptides of unknown identity were found (Extended Data Fig. 4). Notably, this also includes four further gene clusters for proteusins, a recently discovered new natural product family with polytheonamides as the only members known to date[6,24]. Tandom mass spectrometry (MS–MS)-based molecular networking[25] suggested a high diversity of previously unknown metabolite families, indicating that at least some of these orphan pathways are likely to be active (Extended Data Fig. 6). The gene candidates for konbamides

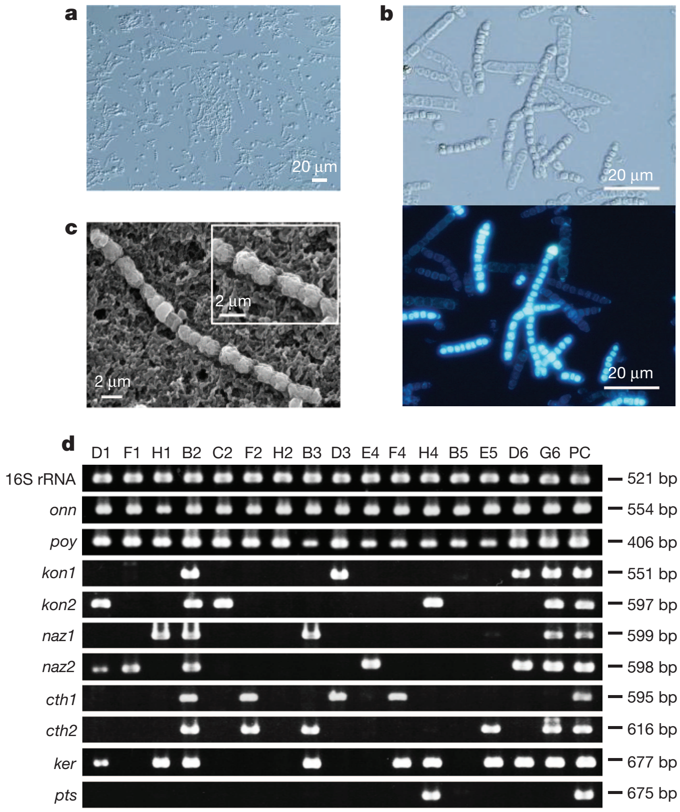


Figure 2 | Single-cell analytic studies. **a**, Differential interference contrast micrograph of the filamentous fraction after differential centrifugation ($n = 3$). **b**, Fluorescence micrograph of filamentous bacteria without (top) and with (bottom) ultraviolet excitation ($n = 3$). **c**, Scanning electron micrograph of a single filamentous bacterium ($n = 3$). **d**, Nested PCR of natural product gene clusters from whole-genome amplification samples of wells sorted with single filaments ($n = 48$). Wells showing positive amplification for '*Entotheonella*' sp. 16S rRNA gene, onnamide (*onn*) and polytheonamide (*poy*) gene clusters ($n = 3$) were used for the identification of the other enzyme clusters ($n = 1$). Each lane represents a single well defined by the well identifier above the top row. *cth*, cyclotheonamide; *ker*, keramamide; *kon*, konbamide; *naz*, nazumamide. PC, metagenomic DNA from filamentous fraction; *pts*, unknown proteusin.

(*kon*), keramamides (*ker*), nazumamide A (*naz*), and an unknown non-ribosomal peptide formed a supercluster of 129 kb. The binning data suggested that this region, the putative cyclotheonamide (*cth*), and the two unassigned proteusin loci all belong to the chromosome of the dominant '*Entotheonella*' sp. TSY1 (Extended Data Fig. 2b). The chromosome of TSY2 contained fewer (at least seven) metabolic gene clusters (two polyketide, at least two NRPS, and a further proteusin cluster) that could not be assigned to known compounds. Except for a small NRPS and a type III polyketide synthase (PKS) system present in both genomes, there was no overlap in the natural product gene repertoires of TSY1 and TSY2, indicating that significant chemical variation exists among members of '*Entotheonella*', even within the same sponge individual. To validate further the source of the plasmid-based polytheonamide and onnamide genes, we conducted additional single-cell experiments (Fig. 2d). All MDA samples previously analysed positive for *onn* and *poy* genes were tested again with PCR primers for various genes of the *kon*, *naz*, *cth*, *ker* and one unknown proteusin pathway. For all cases, positive wells were identified, suggesting that TSY1 carries the plasmid and produces the entire set of metabolites.

Functional evidence for the identity of '*Entotheonella*' gene clusters was obtained by biochemically characterizing gene products from several pathways. Two selected NRPS adenylation domains encoded within the putative *cth* and *ker* pathways were overproduced in *E. coli* and analysed using a γ-$^{18}O_4$-ATP pyrophosphate exchange assay[26] to investigate their amino acid substrate specificity (Extended Data Fig. 7). For the *cth* NRPS, the adenylation domain of module 2 (CthA2) exhibited high selectivity for the rare amino acid 2,3-diaminopropionate (DAP), consistent with the cyclotheonamide structure (Extended Data Fig. 7). The incorporation of this building block is also supported by the presence of two genes in the cluster that encode homologues of SbnAB-type DAP synthases[27]. KerA5 showed greatest substrate specificity for Leu, in agreement with known keramamides and the bioinformatic prediction (Extended Data Fig. 7). Thus, taking the colinearity rule of NRPSs into account, the data support the proposed function of these gene clusters.

We also obtained functional support for a biosynthetic role of the unknown proteusin pathway TSY1_14 by co-expressing the putative nitrile-hydratase-like precursor peptide with a predicted lanthionine synthetase encoded directly adjacent to the precursor gene. Up to three dehydrations of the core peptide were observed by HRMS for the co-expression product compared to the unmodified peptide from expression of the precursor peptide alone. Subsequent alkylation of reduced cysteine residues and tandem MS–MS indicated for each dehydration, one lanthionine bridge was formed within the predicted core peptide (Extended Data Fig. 8 and Supplementary Table 8). These experiments demonstrated that the putative proteusin gene cluster TSY1_14 encodes a functional precursor peptide and modifying lanthionine synthetase. Considering the high complexity of the sponge microbiome, which contains hundreds of ribotypes, the accumulation of metabolic genes in two variants of '*Entotheonella*' is remarkable. Owing to the extraordinary biosynthetic repertoire, we propose the name '*Candidatus* Entotheonella factor' (latin, *factor*; the producer) for these bacteria.

'*Entotheonella*' species are ubiquitous

These findings raised the question whether '*Entotheonella*' spp. also inhabit other sponges and could have a general role in natural product biosynthesis. It was previously shown that an enriched fraction of '*E. palauensis*' from a Palauan chemotype of *T. swinhoei* contained high concentrations of the hybrid polyketide-peptide theopalauamide[14,15]. '*Entotheonella*' members were also detected in another lithistid sponge, *Discodermia dissoluta*, that contains the anticancer polyketide discodermolide[28]. To analyse the distribution of '*Entotheonella*' spp. in depth, 37 taxonomically diverse sponge species collected at 20 locations (Supplementary Table 9) were tested by PCR based on conserved, unique regions of '*Entotheonella*' 16S rRNA genes. Of the 37 sponges, 28 yielded amplicons with sequences exhibiting 95.5–99.9% nucleotide identity to the homologues of '*E. factor*' (Extended Data Fig. 9a, b). Thus, '*Entotheonella*' spp. seem to be widely distributed in marine sponges from distant geographical regions. '*Entotheonella*' amplicons were also obtained from various seawater samples; however, contamination from sponges growing nearby cannot be excluded. For further insights into the discovery potential and chemical variability of these bacteria, we initiated studies on another chemotype of *T. swinhoei* (type W1, referring to the white sponge interior) that contains the actin inhibitor misakinolide A (Extended Data Fig. 10b), a complex polyketide not present in the Y chemotype. PCR detection of PKS genes using the total sponge DNA generated exclusively amplicons that were phylogenetically attributed to the *sup* type (Extended Data Fig. 10c), a putative fatty acid synthase that is widespread and dominant in most sponge microbial consortia and not involved in the production of complex, bioactive polyketides[29]. In contrast, a highly enriched '*Entotheonella*' fraction (Extended Data Fig. 10a) prepared from this sponge yielded a completely different set of amplicons consisting of six gene fragments all belonging to PKSs associated with complex polyketide production (Extended Data Fig. 10c). None of these had a close homologue in TSY1 or TSY2, thus further supporting a diverse chemistry of '*Entotheonella*' phylotypes.

A new candidate phylum, 'Tectomicrobia'

To obtain insights into the taxonomic position of '*Entotheonella*', an initial 16S rRNA-based phylogenetic analysis was conducted (Extended Data Fig. 9c). Altogether, 243 16S rRNA gene sequences were analysed

from marine sponges in this study and from public databases. As the 16S rRNA sequences were only 82% identical to representatives from known bacterial phyla and form a well-separated clade, we suggest the status of a new candidate phylum[30]. The name 'Tectomicrobia' (latin, *tegere*; to hide, to protect) was chosen to reflect their uncultured status as well as the capability to produce bioactive compounds that are likely used as chemical defence. The closest relatives to 'Tectomicrobia' are *Nitrospina* spp., which were recently proposed to belong to a new phylum, Nitrospinae[31]. The known sequences belonging to 'Tectomicrobia' comprise at least three discrete phylogenetic clades. The largest encompasses all '*Entotheonella*' sequences *sensu stricto*, which were largely recovered from marine sponges but also seawater (138 sequences total, of which 107 sequences were produced in this study), a second clade includes related, non-'*Entotheonella*' 16S rRNA gene sequences from various marine sponges (36 sequences), and a third group contains 16S rRNA gene sequences from terrestrial soils (18 sequences). For further validation of the phylogenetic data, we calculated trees using up to 38 concatenated, universally conserved single-copy marker proteins[17] of TSY1, TSY2, and 2,509 bacterial and archaeal taxa to determine the position of '*Entotheonella*' in the tree of life. Recalculations with data sets from closely affiliated phyla (Fig. 3) supported '*Entotheonella*' as belonging to a new sister phylum to Nitrospinae, in agreement with the 16S rRNA data.

Conclusions

Owing to the high frequency of structurally distinct, bioactive metabolites in sponges, these animals have an important role in drug discovery. Compound localization studies suggested Bacteria as producers of individual metabolites[14,15,32,33], but remained ambiguous owing to the possibility of sequestration or transport. The true source of sponge natural products has therefore been a long-standing and, with the exception of metagenomic data providing kingdom-level information[5,6,34], unanswered question. Here we provide evidence that a single member of the highly diverse microbiome of *T. swinhoei* Y, '*E. factor* TSY1', is the source of almost all polyketides and peptides that have been isolated from its sponge host. The bioinformatic assignment to known compounds is further supported by functional studies for polytheonamides[6], onnamide-type compounds[35,36], keramamides, cyclotheonamides and an orphan proteusin. Our data on TSY1, TSY2, and a highly enriched '*Entotheonella*' preparation from a second *T. swinhoei* chemotype, indicate that members of this candidate genus contain producers with a rich and, so far, unique secondary metabolism. Reports on '*Entotheonella*' spp. from two other chemically rich sponges[15,28,37] and our detection of these bacteria in many additional species hint at their more widespread role in the chemistry of their hosts. This study adds the first uncultivated prokaryotes to the taxonomically limited canon of metabolically talented bacteria. '*Entotheonella*' spp. exhibit interesting parallels to streptomycetes and some other well-known producer groups[38–42]; for example, expanded genome size, biosynthetic superclusters[43] and multiple modular assembly lines, high metabolic variability among closely related organisms, and complex morphology. For '*Entotheonella*' spp., complex morphology is particularly noteworthy, as it affords attractive opportunities to systematically study chemical interactions in marine symbioses and to exploit uncultivated bacteria in a targeted way for drug discovery.

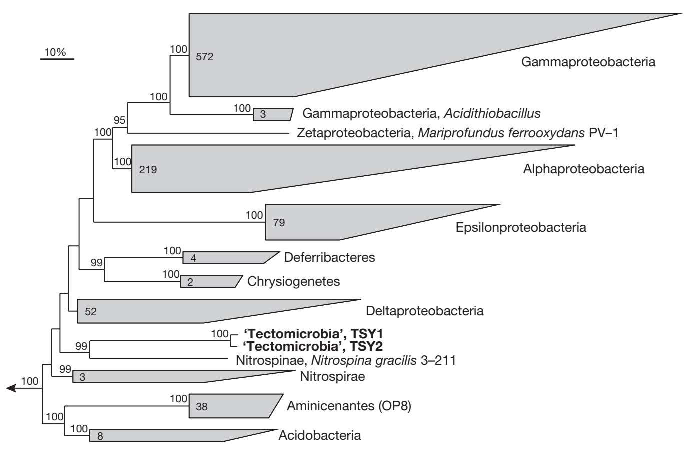


Figure 3 | Phylogenetic inference of the 'Tectomicrobia' and affiliated phyla. RAxML inference of 991 taxa with 100 bootstrap iterations based on up to 38 marker genes. Sequences are collapsed on the phylum level and the number of collapsed sequences is shown for each clade. The two 'Tectomicrobia' variants TSY1 and TSY2 are highlighted in bold. Bootstrap support values of equal or greater than 70% are shown for each node. The scale bar represents 10% estimated sequence divergence. PV-1 and 3-11 are strain names; OP8 is the former name of the (then candidate) phylum Aminicenantes.

METHODS SUMMARY

An adapted differential centrifugation protocol[14] was used to sediment filamentous and unicellular bacteria from the sponge tissue. Single bacteria cells and filaments were sorted into micro-titre plates by flow cytometry with a BD FACSAria II cell sorter (BD Biosciences). Genomic DNA was amplified using an Illustra Genomiphi V2 DNA Amplification Kit (GE Healthcare) and subjected to PCR analysis. Sequence information was obtained using the GS-FLX (454) and MiSeq (Illumina) platforms, using whole-genome sequencing and long mate-pair libraries. Additional sequence reads were obtained by PacBio sequencing (GATC) and Sanger sequencing (IIT). Reads were assembled using the Newbler (v2.6) *de novo* assembler. Automated annotation was performed with Rapid Annotation and Subsystem Technology (RAST)[44] and validated manually. PKS and NRPS domain architecture and substrate specificities were based on sequence alignments and prediction-based software[22,45,46]. Adenylation domains overexpressed in *E. coli* were characterized using a γ-$^{18}O_4$-ATP pyrophosphate exchange assay as previously described[26]. The TSY1_14 proteusin precursor peptide was overexpressed in *E. coli* with and without the putative modifying LanM-like lanthionine synthetase from the same gene cluster. The resulting peptide products were analysed by liquid chromatography (LC)–electrospray ionization (ESI)–HRMS after TCEP (tris-(2-carboxyethyl)-phosphine) treatment, tryptic digest and derivatization. Extracts of *T. swinhoei* and enriched '*Entotheonella*' were analysed by ultra-performance liquid chromatography (UPLC) and nano-LC heated ESI (HESI)–HRMS followed by eMZed[47] data analysis and molecular networking[25].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 June; accepted 18 December 2013.
Published online 29 January 2014.

1. Berdy, J. Bioactive microbial metabolites — a personal view. *J. Antibiot.* **58,** 1–26 (2005).
2. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nature Rev. Microbiol.* **6,** 431–440 (2008).
3. Brady, S. F., Simmons, L., Kim, J. H. & Schmidt, E. W. Metagenomic approaches to natural products from free-living and symbiotic organisms. *Nat. Prod. Rep.* **26,** 1488–1503 (2009).
4. Piel, J. Approaches to capturing and designing biologically active small molecules produced by uncultured microbes. *Annu. Rev. Microbiol.* **65,** 431–453 (2011).
5. Piel, J. *et al.* Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. *Proc. Natl. Acad. Sci. USA* **101,** 16222–16227 (2004).
6. Freeman, M. F. *et al.* Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science* **338,** 387–390 (2012).
7. Hentschel, U. *et al.* Molecular evidence for a uniform microbial community in sponges from different oceans. *Appl. Environ. Microbiol.* **68,** 4431–4440 (2002).
8. Taylor, M. W., Radax, R., Steger, D. & Wagner, M. Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. *Microbiol. Mol. Biol. Rev.* **71,** 295–347 (2007).
9. Hentschel, U., Piel, J., Degnan, S. M. & Taylor, M. W. Genomic insights into the marine sponge microbiome. *Nature Rev. Microbiol.* **10,** 641–654 (2012).
10. Fusetani, N. & Matsunaga, S. Bioactive sponge peptides. *Chem. Rev.* **93,** 1793–1806 (1993).
11. Binga, E. K., Lasken, R. S. & Neufeld, J. D. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J.* **2,** 233–241 (2008).
12. Siegl, A. *et al.* Single-cell genomics reveals the lifestyle of *Poribacteria*, a candidate phylum symbiotically associated with marine sponges. *ISME J.* **5,** 61–70 (2011).
13. Grindberg, R. V. *et al.* Single cell genome amplification accelerates identification of the apratoxin biosynthetic pathway from a complex microbial assemblage. *PLoS ONE* **6,** e18565 (2011).

14. Bewley, C. A., Holland, N. D. & Faulkner, D. J. Two classes of metabolites from *Theonella swinhoei* are localized in distinct populations of bacterial symbionts. *Experientia* **52,** 716–722 (1996).
15. Schmidt, E. W., Obraztsova, A. Y., Davidson, S. K., Faulkner, D. J. & Haygood, M. G. Identification of the antifungal peptide-containing symbiont of the marine sponge *Theonella swinhoei* as a novel δ-proteobacterium, "*Candidatus* Entotheonella palauensis". *Mar. Biol.* **136,** 969–977 (2000).
16. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71,** 3342–3347 (2005).
17. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499,** 431–437 (2013).
18. Chitsaz, H. *et al.* Efficient *de novo* assembly of single-cell bacterial genomes from short-read data sets. *Nature Biotechnol.* **29,** 915–921 (2011).
19. Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic, machinery, and mechanisms. *Chem. Rev.* **106,** 3468–3496 (2006).
20. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6,** 493–505 (1999).
21. Challis, G. L., Ravel, J. & Townsend, C. A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7,** 211–224 (2000).
22. Rottig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39,** W362–W367 (2011).
23. Kobayashi, J. i . *et al.* Konbamide, a novel peptide with calmoduiin antagonistic activity from the Okinawan marine sponge *Theonella* sp. *J. Chem. Soc. Chem. Commun.* 1050–1052 (1991).
24. Haft, D. H., Basu, M. K. & Mitchell, D. A. Expansion of ribosomally produced natural products: a nitrile hydratase-and Nif11-related precursor family. *BMC Biol.* **8,** 70 (2010).
25. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. USA* **109,** E1743–E1752 (2012).
26. Phelan, V. V., Du, Y., McLean, J. A. & Bachmann, B. O. Adenylation enzyme characterization using γ-$^{18}O_4$-ATP pyrophosphate exchange. *Chem. Biol.* **16,** 473–478 (2009).
27. Beasley, F. C., Cheung, J. & Heinrichs, D. E. Mutation of L-2,3-diaminopropionic acid synthase genes blocks staphyloferrin B synthesis in *Staphylococcus aureus. BMC Microbiol.* **11,** 199 (2011).
28. Bruck, W. M., Sennett, S. H., Pomponi, S. A., Willenz, P. & McCarthy, P. J. Identification of the bacterial symbiont *Entotheonella* sp. in the mesohyl of the marine sponge *Discodermia* sp. *ISME J.* **2,** 335–339 (2008).
29. Hochmuth, T. *et al.* Linking chemical and microbial diversity in marine sponges: possible role for poribacteria as producers of methyl-branched fatty acids. *ChemBioChem* **11,** 2572–2578 (2010).
30. Hugenholtz, P., Goebel, B. M. & Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180,** 4765–4774 (1998).
31. Lucker, S., Nowka, B., Rattei, T., Spieck, E. & Daims, H. The genome of *Nitrospina gracilis* illuminates the metabolism and evolution of the major marine nitrite oxidizer. *Front. Microbiol.* **4,** 27 (2013).
32. Unson, M. D., Holland, N. D. & Faulkner, D. J. A brominated secondary metabolite synthesized by the cyanobacterial symbiont of a marine sponge and accumulation of the crystalline metabolite in the sponge tissue. *Mar. Biol.* **119,** 1–11 (1994).
33. Flowers, A. E., Garson, M. J., Webb, R. I., Dumdei, E. J. & Charan, R. D. Cellular origin of chlorinated diketopiperazines in the dictyoceratid sponge *Dysidea herbacea* (Keller). *Cell Tissue Res.* **292,** 597–607 (1998).
34. Fisch, K. M. *et al.* Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nature Chem. Biol.* **5,** 494–501 (2009).
35. Zimmermann, K., Engeser, M., Blunt, J. W., Munro, M. H. & Piel, J. Pederin-type pathways of uncultivated bacterial symbionts: analysis of O-methyltransferases and generation of a biosynthetic hybrid. *J. Am. Chem. Soc.* **131,** 2780–2781 (2009).
36. Pöplau, P., Frank, S., Morinaka, B. I. & Piel, J. An enzymatic domain for cyclic ether formation in complex polyketides. *Angew. Chem. Int. Ed.* **52,** 13215–13218 (2013).
37. Schirmer, A. *et al.* Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta. Appl. Environ. Microbiol.* **71,** 4840–4849 (2005).
38. Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417,** 141–147 (2002).
39. Omura, S. *et al.* Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. *Proc. Natl. Acad. Sci. USA* **98,** 12215–12220 (2001).
40. Schneiker, S. *et al.* Complete genome sequence of the myxobacterium *Sorangium cellulosum. Nature Biotechnol.* **25,** 1281–1289 (2007).
41. Frangeul, L. *et al.* Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* **9,** 274 (2008).
42. Flores, E. & Herrero, A. Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nature Rev. Microbiol.* **8,** 39–50 (2010).
43. Mast, Y. *et al.* Characterization of the 'pristinamycin supercluster' of *Streptomyces pristinaespiralis. Microb. Biotechnol.* **4,** 192–206 (2011).
44. Aziz, R. K. *et al.* The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9,** 75–89 (2008).
45. Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39,** W339–W346 (2011).
46. Bachmann, B. O. & Ravel, J. Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol.* **458,** 181–217 (2009).
47. Kiefer, P., Schmitt, U. & Vorholt, J. A. eMZed: an open source framework in Python for rapid and interactive development of LC/MS data analysis workflows. *Bioinformatics* **29,** 963–964 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Lasken and M. F. Freeman for discussion, R. W. M. van Soest, P. R. Bergquist, and Y. Ise for taxonomic identification of sponges, P. Kiefer for eMZed support, P. Dorrestein and J. Watrous for mass spectrometry networking support, A. Semeniuk and R. Meoded for experimental support, and T. Ravasi, P. Crews, Y. Kashman and M. Aknin for providing sponge specimens. This work was supported by the SNF (31003A_146992) to J.P., BMBF (GenBioCom: 0315581I to J.P. and 0315585J to J.K.), DFG (PI 430/1-3 and PI 430/9-1 to J.P., SFB 630-TP A5 to U.H.), the EU (BlueGenics to J.P.), MIWFT within the BIO.NRW initiative (280371902 to C. Rückert), the Grants-in-aid for Young Scientists (B), KAKENHI (23760755 to T.M.), JSPS to J.P., S.M. and H.T., Alexander von Humboldt Foundation to M.C.W., German National Academic Foundation to M.J.H. and E.J.N.H., and DAAD to A.R.U. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract no. DE-AC02-05CH11231.

Author Contributions J.P., H.T., T.M. and J.K. conceived the study. S.M., T.W., K.T., I.A. and U.H. collected the sponges and determined the chemotypes. J.P. performed the cell separation and differential centrifugation, C.G. and A.R.U. isolated the DNA, T.M. and E.J.N.H. conducted the single-cell studies. C.Rü. and J.K. sequenced the metagenome, M.C.W., C. Rü., J.P., U.A.E.S., N.H., C.G., A.R.U. and M.J.H. analysed genomic data, K.T. and C.G. performed the distribution studies. M.C., M.K., and M.C.W. performed the adenylation domain assays, M.J.H. performed the proteusin studies, A.R.U. performed the studies on the misakinolide chemotype, C. Ri. and S.S. performed the phylogenetic analysis, and A.O.B. and E.J.N.H. performed HRMS experiments. All authors planned the experiments, analysed the data and wrote the manuscript.

Author Information Sequence data for 16S rRNA have been deposited in GenBank under accession numbers KF926701–KF926822. Sequence data for Whole Genome Shotgun projects have been deposited at DDBJ/EMBL/GenBank under the accession AZHW00000000 and AZHX00000000. The versions described in this paper are AZHW01000000 and AZHXW01000000, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.P. (jpiel@ethz.ch).

LETTER

doi:10.1038/nature12898

A cosmic web filament revealed in Lyman-α emission around a luminous high-redshift quasar

Sebastiano Cantalupo[1,2], Fabrizio Arrigoni-Battaia[3], J. Xavier Prochaska[1,2,3], Joseph F. Hennawi[3] & Piero Madau[1]

Simulations of structure formation in the Universe predict that galaxies are embedded in a 'cosmic web'[1], where most baryons reside as rarefied and highly ionized gas[2]. This material has been studied for decades in absorption against background sources[3], but the sparseness of these inherently one-dimensional probes preclude direct constraints on the three-dimensional morphology of the underlying web. Here we report observations of a cosmic web filament in Lyman-α emission, discovered during a survey for cosmic gas fluorescently illuminated by bright quasars[4,5] at redshift $z \approx 2.3$. With a linear projected size of approximately 460 physical kiloparsecs, the Lyman-α emission surrounding the radio-quiet quasar UM 287 extends well beyond the virial radius of any plausible associated dark-matter halo and therefore traces intergalactic gas. The estimated cold gas mass of the filament from the observed emission—about $10^{12.0 \pm 0.5}/C^{1/2}$ solar masses, where C is the gas clumping factor—is more than ten times larger than what is typically found in cosmological simulations[5,6], suggesting that a population of intergalactic gas clumps with subkiloparsec sizes may be missing in current numerical models.

A recent pilot survey[5] using a custom-built, narrow-band filter on the Very Large Telescope demonstrated that bright quasars can, like a flashlight, 'illuminate' the densest knots in the surrounding cosmic web and boost fluorescent Lyman-α emission[4,5,7–9] to detectable levels. Following the same experiment, we imaged UM 287 on 2012 November 12 and 13 UT with a custom narrow-band filter (NB3985) tuned to Lyman α at $z = 2.28$ inserted into the camera of the Low Resolution Imaging Spectrometer (LRIS) on the 10-m Keck I telescope (see Extended Data Fig. 1). Figure 1 presents the processed and combined images, centred on UM 287. In the NB3985 image, we identify a very extended nebula originating near the quasar with a projected size of about 1 arcmin. In the broad-band images no extended emission is observed. This requires the narrow-band light to be line-emission, and we identify it as Lyman α at the redshift of UM 287.

Figure 2 presents the NB3985 image, continuum subtracted using standard techniques (see Methods) and smoothed with a 1-arcsec Gaussian kernel. This image is dominated by the filamentary and asymmetric nebula that has a maximum projected extent of 55 arcsec as defined by the 10^{-18} erg s^{-1} cm^{-2} arcsec^{-2} isophotal contour, corresponding to about 460 physical kpc or 1.5 Mpc in co-moving coordinates. Including (excluding) the emission from UM 287 falling within the narrow-band filter, the structure has a total line luminosity $L_{\mathrm{Ly}\alpha} = (1.43 \pm 0.05) \times 10^{45}$ erg s^{-1} ($L_{\mathrm{Ly}\alpha} = (2.2 \pm 0.2) \times 10^{44}$ erg s^{-1}).

Although Lyman-α nebulae extending up to about 250 kpc have been previously detected[10–14], the UM 287 nebula represents a system that is unique so far: given its size, it extends well beyond any plausible dark-matter halo associated with UM 287 (see below), representing an exceptional example of emitting gas on intergalactic scales.

The largest Lyman-α nebulae previously discovered (see Fig. 3) are associated with the most massive dark-matter haloes present in the high-redshift Universe. High-redshift radio galaxies (HzRGs), inferred to host obscured but luminous active galactic nuclei (AGN)[11,13], are often surrounded by giant Lyman-α envelopes extending up to about 250 kpc at $z \approx 3$ (ref. 15). Clustering arguments and the observation of large overdensities of Lyman-α galaxies, together with the lack of X-ray detection from a possible intracluster medium, suggest that HzRGs are associated with haloes of 10^{13} solar masses ($M_\odot$)[15–17]. With a virial diameter of about 300 kpc at $z \approx 3$, these haloes are therefore able to contain the largest HzRG Lyman-α nebulae. Blind narrow-band surveys have derived an apparently different population of large nebulae (termed Lyman-α blobs) with sizes extending up to 180 physical kpc at $z \approx 3$ that, in some cases, do not appear to be associated with a particular bright galaxy or AGN[12,14,18,19]. The rarity and the strong clustering of these sources, suggest, as for HzRGs, an association with proto-cluster environments and haloes with masses of about $10^{13} M_\odot$ (refs 20, 21). Although the detailed origin of the emission of the Lyman-α blobs is still unclear, the sizes of the associated haloes strongly suggest that the emitting gas is confined within the halo itself. This is also the case for the Lyman-α nebulae previously detected around a small number of bright quasars, extending up to about 100 kpc (refs 10, 22–24). Clustering studies demonstrate that bright quasars at $z < 3$ populate haloes of mass $\sim 10^{12.5} M_\odot$ (that have a virial diameter of about 280 kpc at $z \approx 2.3$) independently of their redshift or luminosity[25,26].

The exceptional nature of the nebula is due not only to its size (about 460 physical kpc) but also to the fact that it is associated with a radio-quiet quasar. Radio-quiet quasars have the smallest host halo mass ($\sim 10^{12.5} M_\odot$) and virial diameter (280 kpc) among previously detected objects and do not have radio-emitting jets that may power Lyman-α emission on large scales[27]. In order for the nebula to be fully contained within the virial radius of a dark-matter halo centred on UM 287, a halo mass would be required that is at least ten times larger than the typical value associated with radio-quiet quasars. This would make the host halo of UM 287 one of the largest known at $z > 2$, a possibility that is excluded by the absence of a significant overdensity of Lyman-α emitters around UM 287 compared to other radio-quiet quasars (see Methods). Differently from any previous detection, the nebula is therefore an image of intergalactic gas at $z > 2$ extending beyond any individual, associated dark-matter halo. The rarity of these systems may be explained by the combination of anisotropic emission from the quasars (typically only about 40% of the solid angle around a bright, high-redshift quasar is unobstructed[28]), the anisotropic distribution of dense filaments and light travel effects that, for quasar ages of less than a few million years, further limit the possible 'illuminated' volume.

In order to constrain the physical properties of this system, we use a set of Lyman-α radiative transfer calculations[29] combined with adaptive mesh refinement simulation of cosmological structure formation around a dark-matter halo with mass $M_{\mathrm{DM}} \approx 10^{12.5} M_\odot$ (see Methods). We consider two possible, extreme scenarios for the Lyman-α emission mechanism of the intergalactic gas associated with the nebula: (1) the gas is highly ionized by the quasar and the Lyman-α emission is mainly produced by hydrogen recombinations; and (2) the gas is mostly neutral and the emission is mainly due to scattering of the Lyman-α and continuum photons produced by the quasar broad line region. The

[1]Department of Astronomy and Astrophysics, University of California, 1156 High Street, Santa Cruz, California 95064, USA. [2]University of California Observatories, Lick Observatory, 1156 High Street, Santa Cruz, California 95064, USA. [3]Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany.

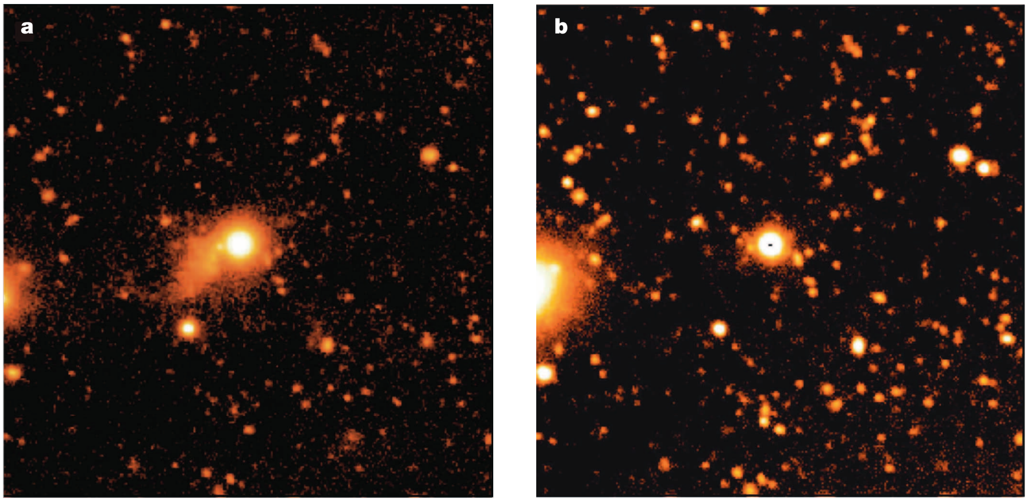


Figure 1 | Processed and combined images of the field surrounding the quasar UM 287. a, b, Each image is 2 arcmin on a side, and the quasar is located at the centre. In the narrow-band (NB3985) image (**a**), which is tuned to the Lyman-α line of the systemic redshift for UM 287, we identify very extended (~55 arcsec across) emission. The deep V-band image (**b**) does not show any extended emission associated with UM 287. This requires the nebula to be line-emission, and we identify it as Lyman-α at the redshift of the quasar.

models are used to obtain scaling relations between the observable Lyman-α surface brightness from the intergalactic gas surrounding the quasar and the hydrogen column densities (see Extended Data Fig. 3). These scaling relations are consistent with analytical expectations. Note that the estimated column densities for scenario (1) depend on the ionized gas clumping factor ($C = \langle n_e^2 \rangle / \langle n_e \rangle^2$, where n_e is the electron density) below the simulation resolution scale, ranging from about 10 physical kpc for diffuse intergalactic gas to ~160 physical pc for the densest regions within galaxies.

The results are presented in Fig. 4. The observed Lyman-α emission requires very large column densities of 'cold' ($T < 5 \times 10^4$ K) gas, up to $N_H \approx 10^{22}$ cm^{-2}. The implied total, cold gas mass 'illuminated' by

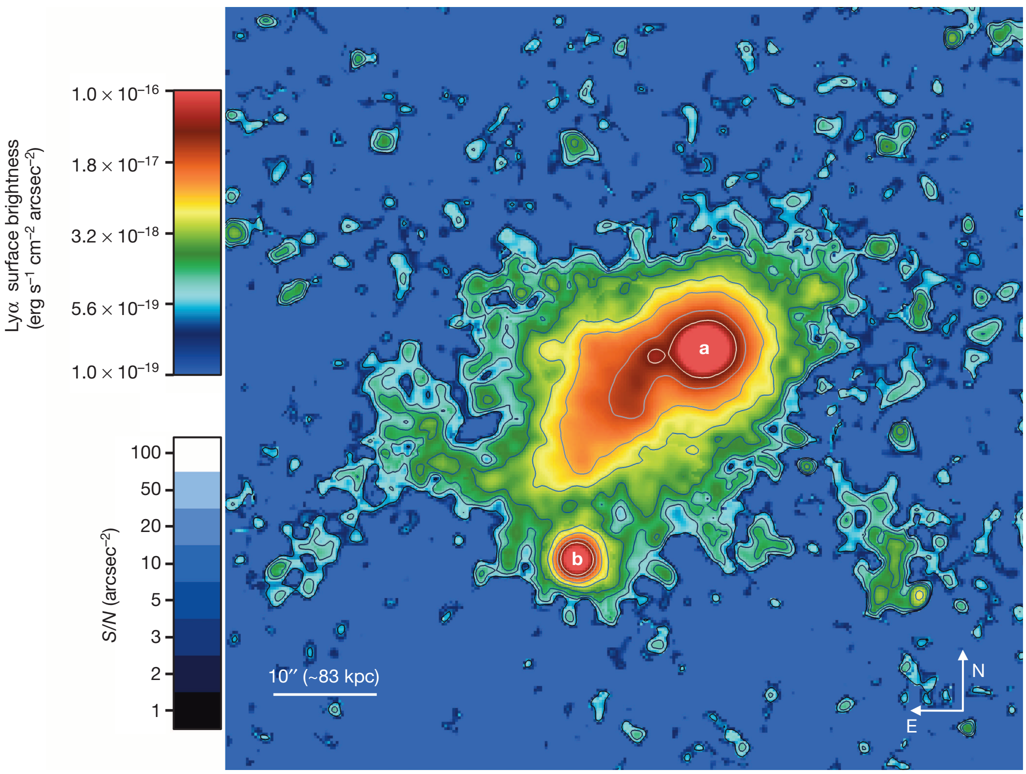


Figure 2 | Lyman-α image of the UM 287 nebula. We subtracted from the narrow-band image the continuum contribution estimated from the broad-band images (see Methods). The location of UM 287 is labelled with 'a'. The colour map and the contours indicates, respectively, the Lyman-α (Lyα) surface brightness (upper colour scale) and the signal-to-noise ratio per arcsec^2 aperture (lower colour scale). The extended emission spans a projected angular size of ~55 arcsec (about 460 physical kpc), measured from the 2σ (~10^{-18} erg s^{-1} cm^{-2} arcsec^{-2}) contours. The object marked with 'b' is an optically faint ($g \approx 23$AB) quasar at the same redshift as UM 287 (see Extended Data Fig. 2). The nebula appears broadly filamentary and asymmetric, extending mostly on the eastern side of quasar UM 287 up to a projected distance of about 35 arcsec (~285 physical kpc) measured from the 2σ isophotal. The nebula extends towards the southeast in the direction of the optically faint quasar. However, the two quasars do not seem to be directly connected by this structure that continues as a fainter and spatially narrower filament. The large distance between the two quasars and the very broad morphology of the nebula argue against the possibility that it may originate from an interaction between the quasar host galaxies (see Methods).

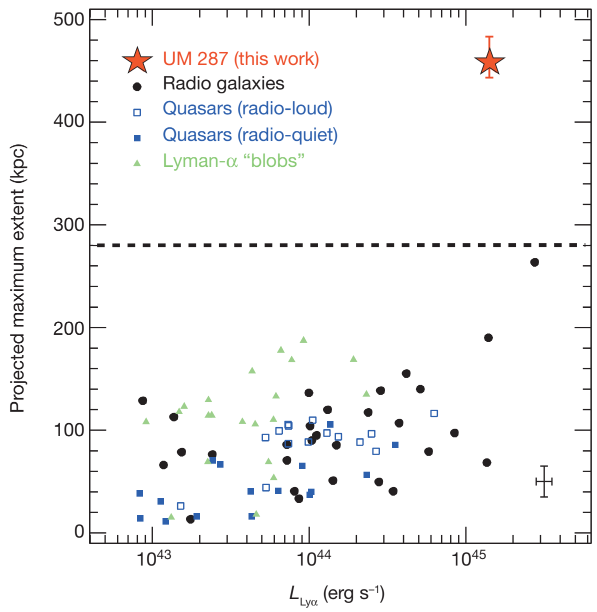


Figure 3 | Luminosity–size relations for previously detected, bright Lyman-α nebulae and UM 287. The plot includes nebulae surrounding radio galaxies (black circles), radio-loud quasars (blue open squares), radio-quiet quasars (blue filled squares) and Lyman-α 'blobs' (green triangles). The reported luminosities include the Lyman-α ($L_{Ly\alpha}$) emission (within the narrow-band filters) from any sources embedded in the nebulae, if present. Excluding the contribution coming directly from the quasar broad line region, the luminosity of the UM 287 nebula corresponds to $L_{Ly\alpha} = 2.2 \pm 0.2 \times 10^{44}$ erg s^{-1} (about 16% of the total luminosity). Error bars for UM 287 represent the 1σ photometric error including continuum-subtraction (error bar is smaller than the symbol size) and an estimate of the error on the projected maximum extent using $\pm 1\sigma$ isophotal contours with respect to the 10^{-18} erg s^{-1} cm^{-2} arcsec^{-2} isophotal. The typical errors for other sources are presented separately in the bottom-right corner. The dashed line indicates the virial diameter of a dark-matter halo with total mass $M \approx 10^{12.5} M_\odot$, the typical host of radio-quiet quasars including UM 287, as confirmed by the analysis of the galaxy overdensity in our field (see Methods). The UM 287 nebula, differently from any previous detection, extends on intergalactic medium scales that are well beyond any possible associated dark-matter halo. Note that even if we restrict the size measurement of the UM 287 nebula to the 4×10^{-18} erg s^{-1} cm^{-2} arcsec^{-2} isophotal to be comparable with the majority of the previous surveys, the measured apparent size of the UM 287 nebula will be reduced only by about 20%.

the quasar is $M_{gas} \approx 10^{12 \pm 0.5} M_\odot$ for the 'mostly ionized' case (scenario (1)) assuming $C = 1$ and $M_{gas} \approx 10^{11.4 \pm 0.6} M_\odot$ for the 'mostly neutral' case (scenario (2)). Note that the total estimated mass for case (1) scales as $C^{-1/2}$. For comparison, a typical simulated filament in our cosmological simulation of structure formation with size and morphology similar to the nebula around a dark-matter halo of mass $M_{DM} \approx 10^{12.5} M_\odot$ has a total gas mass of about $M_{gas} \approx 10^{11.3} M_\odot$, but only about 15% of this gas is 'cold' ($T < 5 \times 10^4$ K)—that is, $M_{gas} \approx 10^{10.5} M_\odot$—and therefore able to emit substantial Lyman-α emission. These estimates are consistent with a large sample of simulated haloes obtained by other recent works based on cosmological adaptive mesh refinement simulations[6]. These simulations also show a (weak) decreasing trend of the cold gas fraction with halo mass.

How can we explain the large differences between the estimated mass of cold gas in the nebula and the available amount of cold gas predicted by numerical simulations on similar scales? One possibility is to assume that the simulations are not resolving a large population of small, cold gas clumps within the low-density intergalactic medium that are illuminated and ionized by the intense radiation of the quasar. In this case, an extremely high clumping factor, up to $C \approx 1{,}000$, on scales below a few kiloparsecs would be required in order to explain the large luminosity of the nebula with the cold gas mass predicted by the

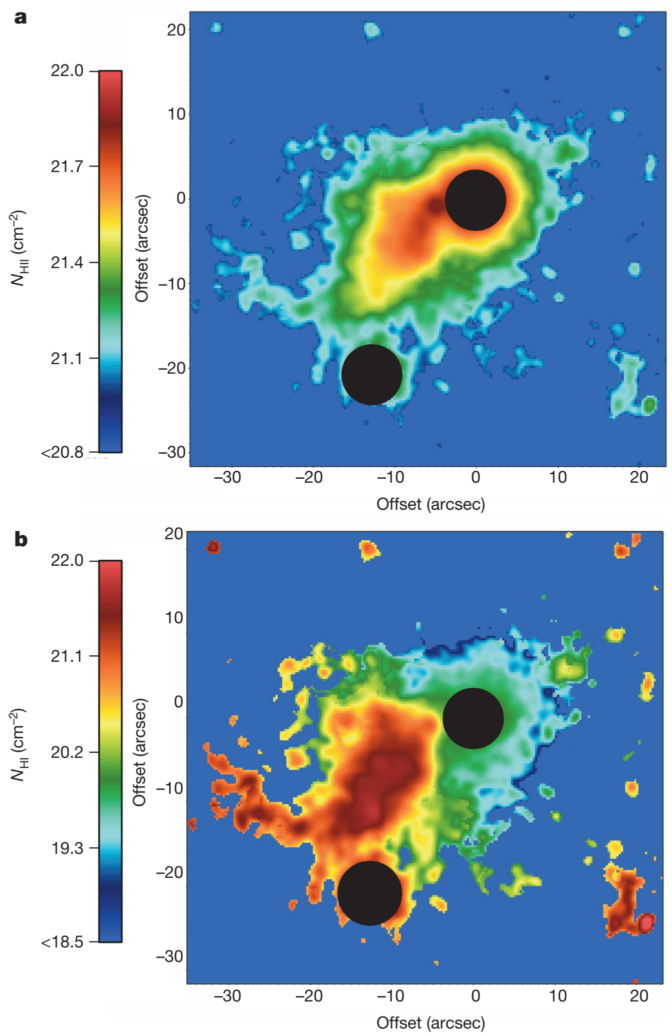


Figure 4 | Inferred hydrogen column densities associated with the UM 287 nebula. We have converted the observed Lyman-α surface brightness into gas column densities N using a set of scaling relations obtained with detailed radiative transfer simulations and consistent with analytical expectations (see Extended Data Fig. 3 and Methods). We have explored two extreme cases: first, the gas is mostly ionized by the quasar radiation (**a**; N_{HII}) and second, the gas is mostly neutral (**b**; N_{HI}). Two circular regions with a diameter of 7 arcsec (~8 times the seeing radius) have been masked at the location of the quasars (black circles). The inferred hydrogen column density in **a** scales as $C^{-1/2}$, where C is the gas clumping factor on a spatial scale of about 10 physical kpc at moderate overdensities (less than about 40 times the mean density of the Universe at $z = 2.28$). The implied column densities and gas masses, in both cases, are at least a factor of ten larger than what is typically observed within cosmological simulations around massive haloes, suggesting that a large number of small clumps within the diffuse intergalactic medium may be missing within current numerical models.

simulations. On the other hand, if some physical process that is not fully captured by current grid-based simulations increases the fraction of cold gas around the quasar—for example, a proper treatment of metal mixing—a smaller clumping factor may be required. In the extreme (and rather unrealistic) case that all the hot gas is turned into a cold phase, the required clumping factor would be $C \approx 20$. Even if the gas is not ionized by the quasar (scenario (2) above), the simulations are able to reproduce the observed mass only if a substantial amount of hot gas is converted into a cold phase. Incidentally, this is exactly the same result produced by comparing the properties of Lyman-α absorption systems around a large statistical sample of quasars with simulations[30]. Proper modelling of this gas phase will require a new generation of

numerical models that are able—simultaneously—to spatially resolve these small intergalactic clumps within large simulation boxes, and to treat the multiphase nature of this gas and its interaction with galaxies and quasars.

METHODS SUMMARY

We observed UM 287 for a total of 10 h in a series of dithered, 1,200-s exposures. In parallel, we obtained 10 h of broad-band V images with the LRIS-red camera and 1 h of B-band imaging. For all observations, we used the D460 dichroic beam splitter. We binned the blue CCDs 2×2 to minimize read noise. The images were processed using standard routines within the reduction software IRAF, including bias subtraction, flat fielding and illumination correction. A combination of twilight sky flats and unregistered science frames has been used to produce flat-field images and illumination corrections for each band. We have calibrated the photometry of our images using two spectrophotometric stars (Feige 110 and Feige 34) and the standard star field PG 0231+051. To isolate the emission in the Lyman-α line we estimated and then subtracted the continuum emission from discrete and extended sources contained within the NB3985 filter using a combination of the V band and B band. We derived a relation between the observable Lyman-α emission from diffuse gas illuminated by a quasar and the gas column densities by combining a Lyman-α radiative transfer model with the results of a cosmological hydrodynamical simulation of structure formation at $z = 2.3$ (ref. 5). The cosmological simulation consists of a 40^3 co-moving Mpc^3 cosmological volume with a 10^3 co-moving Mpc^3 high-resolution region containing a massive halo compatible with the expected quasar hosts ($M_{\mathrm{DM}} \approx 10^{12.5} M_{\odot}$). The equivalent base-grid resolution in the high-resolution region corresponds to a $(1{,}024^3)$ grid with a dark-matter particle mass of about $1.8 \times 10^6 M_{\odot}$. We adaptively refined the grid by a factor of 2^6, reaching a maximum spatial resolution of about 0.6 co-moving kpc, that is, about 165 proper pc at $z = 2.3$.We have then applied in post processing an ionization and Lyman-α radiative transfer using the RADAMESH adaptive mesh refinement code[29].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 February; accepted 20 November 2013.
Published online 19 January 2014.

1. Bond, J. R., Kofman, L. & Pogosyan, D. How filaments of galaxies are woven into the cosmic web. *Nature* **380,** 603–606 (1996).
2. Cen, R., Miralda-Escudé, J., Ostriker, J. P. & Rauch, M. Gravitational collapse of small-scale structure as the origin of the Lyman-alpha forest. *Astrophys. J.* **437,** L9–L12 (1994).
3. Rauch, M. The Lyman alpha forest in the spectra of QSOs. *Annu. Rev. Astron. Astrophys.* **36,** 267–316 (1998).
4. Cantalupo, S., Porciani, C., Lilly, S. J. & Miniati, F. Fluorescent Lyα emission from the high-redshift intergalactic medium. *Astrophys. J.* **628,** 61–75 (2005).
5. Cantalupo, S., Lilly, S. J. & Haehnelt, M. G. Detection of dark galaxies and circum-galactic filaments fluorescently illuminated by a quasar at z = 2.4. *Mon. Not. R. Astron. Soc.* **425,** 1992–2014 (2012).
6. Fumagalli, M. *et al.* Confronting simulations of optically thick gas in massive halos with observations at z=2–3. Preprint at http://arXiv.org/abs/1308.1669 (2013).
7. Gould, A. & Weinberg, D. H. Imaging the forest of Lyman limit systems. *Astrophys. J.* **468,** 462–468 (1996).
8. Cantalupo, S., Lilly, S. J. & Porciani, C. Plausible fluorescent Lyα emitters around the z = 3.1 QSO 0420–388. *Astrophys. J.* **657,** 135–144 (2007).
9. Kollmeier, J. A. *et al.* Lyα emission from cosmic structure. I. Fluorescence. *Astrophys. J.* **708,** 1048–1075 (2010).
10. Heckman, T. M., Miley, G. K., Lehnert, M. D. & van Breugel, W. Spatially resolved optical images of high-redshift quasi-stellar objects. *Astrophys. J.* **370,** 78–101 (1991).
11. McCarthy, P. J. High redshift radio galaxies. *Annu. Rev. Astron. Astrophys.* **31,** 639–688 (1993).
12. Steidel, C. C. *et al.* Lyα imaging of a proto-cluster region at <z>=3.09. *Astrophys. J.* **532,** 170–182 (2000).
13. Reuland, M. *et al.* Giant Lyα nebulae associated with high-redshift radio galaxies. *Astrophys. J.* **592,** 755–766 (2003).
14. Matsuda, Y. *et al.* The Subaru Lyα blob survey: a sample of 100-kpc Lyα blobs at z = 3. *Mon. Not. R. Astron. Soc.* **410,** L13–L17 (2011).
15. Venemans, B. P. *et al.* Protoclusters associated with z > 2 radio galaxies. I. Characteristics of high redshift protoclusters. *Astron. Astrophys.* **461,** 823–845 (2007).
16. Hickox, R. C. *et al.* Host galaxies, clustering, Eddington ratios, and evolution of radio, X-ray, and infrared-selected AGNs. *Astrophys. J.* **696,** 891–919 (2009).
17. Carilli, C. L. *et al.* The X-ray-radio alignment in the z=2.2 radio galaxy PKS 1138–262. *Astrophys. J.* **567,** 781–789 (2002).
18. Dey, A. *et al.* Discovery of a large ~200 kpc gaseous nebula at z ~ 2.7 with the Spitzer Space Telescope. *Astrophys. J.* **629,** 654–666 (2005).
19. Prescott, M. K. M., Dey, A. & Jannuzi, B. T. The discovery of a large Lyα+He II nebula at z ~ 1.67: a candidate low metallicity region? *Astrophys. J.* **702,** 554–566 (2009).
20. Yang, Y., Zabludoff, A., Eisenstein, D. & Davé, R. Strong field-to-field variation of Lyα nebulae populations at z ~ 2.3. *Astrophys. J.* **719,** 1654–1671 (2010).
21. Geach, J. E. *et al.* The Chandra deep protocluster survey: Lyα blobs are powered by heating, not cooling. *Astrophys. J.* **700,** 1–9 (2009).
22. Bergeron, J. *et al.* Lyα emission at $Z \sim z_{\mathrm{em}}$ around the quasar J2233–606 in the Hubble Deep Field South. *Astrophys. J.* **343,** L40–L44 (1999).
23. Christensen, L., Jahnke, K., Wisotzki, L. & Sánchez, S. F. Extended Lyman-α emission around bright quasars. *Astron. Astrophys.* **459,** 717–729 (2006).
24. North, P. L., Courbin, F., Eigenbrod, A. & Chelouche, D. Spectroscopy of extended Lyα envelopes around z = 4.5 quasars. *Astron. Astrophys.* **542,** A91 (2012).
25. DaÂngela, J. *et al.* The 2dF-SDSS LRG and QSO survey: QSO clustering and the L-z degeneracy. *Mon. Not. R. Astron. Soc.* **383,** 565–580 (2008).
26. Trainor, R. F. & Steidel, C. C. The halo masses and galaxy environments of hyperluminous QSOs at z ≅ 2.7 in the Keck baryonic structure survey. *Astrophys. J.* **752,** 39 (2012).
27. Villar-Martín, M. *et al.* VIMOS-VLT spectroscopy of the giant Lyα nebulae associated with three z ~ 2.5 radio galaxies. *Mon. Not. R. Astron. Soc.* **378,** 416–428 (2007).
28. Polletta, M. *et al.* Obscuration in extremely luminous quasars. *Astrophys. J.* **675,** 960–984 (2008).
29. Cantalupo, S. & Porciani, C. RADAMESH: cosmological radiative transfer for adaptive mesh refinement simulations. *Mon. Not. R. Astron. Soc.* **411,** 1678–1694 (2011).
30. Prochaska, J. X. *et al.* Quasars probing quasars VI. Excess HI absorption within one proper Mpc of z~2 quasars. *Astrophys. J.* **776,** 136 (2013).

Acknowledgements We thank the staff of the W.M. Keck Observatory for their support during the installation and testing of our custom-built narrow-band filter. S.C. thanks M. Haehnelt for comments on an earlier version of the letter and J. Primack for useful conversations. S.C. and J.X.P. acknowledge support from the National Science Foundation (NSF) grant AST-1010004. P.M. acknowledges support from the NSF through grant OIA-1124453, and from NASA through grant NNX12AF87G. The data presented here were obtained at the W.M. Keck Observatory, which is operated as a scientific partnership among the California Institute of Technology, the University of California and NASA. The Observatory was made possible by the financial support of the W.M. Keck Foundation. We acknowledge the cultural role that the summit of Mauna Kea has within the indigenous Hawaiian community.

Author Contributions S.C. designed the observational survey and the custom-built filter, conducted the observations, led the narrow-band imaging data reduction and analysis, performed the numerical simulations and led the theoretical interpretation, the writing of the text and the production of the figures. F.A.-B. and J.X.P. assisted with the observations, contributed to data reduction, the text and the figures. In particular, F.A.-B. reduced and calibrated the images, produced the continuum-subtracted image, the catalogues of Lyman-α emitters, and compiled data on all Lyman-α nebulae in the literature. J.X.P. reduced the spectrum of the companion quasar and contributed to the text. J.F.H. and P.M. contributed to the text and assisted with the planning and interpretation of the observations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.C. (cantal@ucolick.org).

METHODS

Observations and data reduction. As part of a continuing programme to search for Lyman-α emission associated with the fluorescence of quasar ionizing radiation[5], we obtained deep, narrow-band imaging of the field surrounding UM 287, also known as PHL 868 and LBQS 0049+0045. UM 287 was discovered in the University of Michigan emission-line survey, has a precisely measured redshift $z = 2.279 \pm 0.001$ based on analysis of [O III] emission lines[31], and has a bolometric luminosity $L_{\rm bol} \approx 10^{47.3}\,{\rm erg\,s^{-1}}$ estimated from its 1,450-Å rest-frame flux using standard cosmology[32]. This places it in the upper quartile of ultraviolet-bright quasars at this redshift. Assuming that the spectral energy distribution follows a power law[33] with frequency index $\alpha = -1.57$ at energies exceeding 1 R, we estimate the luminosity of ionizing photons[34] to be $\Phi = 10^{57.3}\,{\rm s^{-1}}$ assuming isotropic emission.

The quasar has no counterpart in the FIRST[35] images at 20 cm (1.4 GHz), and based on the FIRST coverage maps we obtain a 5σ flux limit $F_{\rm radio} < 0.76$ mJy, which, given its large ultraviolet luminosity, classifies this quasar as radio-quiet[36]. We selected this source for imaging based solely on its high luminosity, its precisely measured redshift, and its radio-quiet characteristics. We purchased a custom-designed narrow-band filter from Andover Corporation, sized to fit within the grism holder of the Keck/LRISb camera. The filter was tuned to Lyman α at the source's systemic redshift and we requested a narrow band-pass (full-width at half-maximum FWHM ≈ 3 nm) that minimized sky background while maximizing throughput. Extended Data Fig. 1 presents the as-measured transmission curve of the NB 3985 filter.

We observed UM 287 on the nights of UT 12–13 November 2013 for a total of 10 h, in a series of dithered, 1,200 s exposures. Conditions were clear, with atmospheric seeing varying from FWHM ≈ 0.6–1 arcsec. In parallel, we obtained 10 h of broad-band V images with the LRISr camera and 1 h of B-band imaging. For all observations, we employed the D460 dichroic beam splitter. We binned the blue CCDs 2 × 2 to minimize read noise.

All of these data were processed with standard techniques. Bias subtraction was performed using measurements from the overscan regions of each image. The images have been reduced using standard routines within the reduction software IRAF, including bias subtraction, flat fielding and illumination correction. A combination of twilight sky flats and unregistered science frames has been used to produce flat-field images and illumination corrections for each band. Each individual frame has been registered on the SDSS-DR7 catalogue using SExtractor[37] and SCAMP[38] in sequence. The astrometric uncertainty of our registered images is about 0.2 arcsec. Finally, for each band (NB3985, B, V), the corrected frames were average-combined using SWarp[39].

We have calibrated the photometry of our images in the following manner. First, we observed during the two nights two spectrophotometric stars (Feige 110 and Feige 34) through the narrow-band filter, under clear conditions. For the broad band images, we observed the standard star field PG 0231+051.

To compute the zero-point for the narrow-band images, we first measured the number of counts per second of the standard stars Feige 110 and Feige 34. We then compared this measurement with the flux expected, estimated by convolving the spectrum of the standard star with the normalized filter transmission curve (Extended Data Fig. 1). The two measurements agreed to within 0.1 mag. We attribute the difference to small variations in the transparency and adopt an average zero-point of 24.14 mag. The surface brightness limit for our observation in the central region of the image occupied by the nebula is about $5 \times 10^{-19}\,{\rm erg\,s^{-1}\,cm^{-2}\,arcsec^{-2}}$ at 1σ level within an aperture of 1 arcsec^2.

For the broad-band images, we compared the number of counts per second of the five stars in the PG 0231+051 field with their tabulated V and B magnitudes[40]. The derived zero-point for the five stars are consistent with each other within a few percent and we adopt the average values: $B_{\rm ZP} = 28.40$ mag and $V_{\rm ZP} = 28.07$ mag.

As the standard stars and the PG 0231+051 field were observed with a similar airmass of approximately 1.2, which corresponds to the average airmass of our observations, we did not correct the individual images before combination. Moreover, by monitoring unsaturated stars on several exposures, we estimated that the correction would be of the order of a few percent.

Continuum subtraction. To isolate the emission in the Lyman-α, line we estimated and then subtracted the continuum emission from discrete and extended sources contained within the NB3985 filter. We estimated the continuum using a combination of the V-band and B-band images as follows. First, we smoothed both of the broad-band images using a Gaussian kernel of 1 arcsec and set to zero all of the pixels with values less than the measured root-mean-square (1σ). Additionally, in the V-band we set to zero all of the pixels which have signal above 1σ in the B band, as we prefer to use the latter image when possible given that it lies closer in wavelength to the Lyman-α line.

After matching the seeing between the narrow-band and the broad-band images, the continuum subtraction has been applied using the following formula

$$\mathrm{Ly}\alpha = \mathrm{NB3985} - a\left(\frac{\mathrm{FWHM_{NB3985}}}{\mathrm{FWHM_B}}\right)\left(\frac{\mathrm{Tr_{NB3985}}}{\mathrm{Tr_B}}\right)B - b\left(\frac{\mathrm{FWHM_{NB3985}}}{\mathrm{FWHM_V}}\right)\left(\frac{\mathrm{Tr_{NB3985}}}{\mathrm{Tr_V}}\right)V$$

where Lyα is the final subtracted image, NB3985 is the smoothed narrow-band image, B and V are the smoothed and masked broad-band images, and $\mathrm{Tr_{NB3985}}$, $\mathrm{Tr_B}$ and $\mathrm{Tr_V}$ are the transmission peak values for NB3985, B-band and V-band filters, respectively. The parameters $a = 0.85$ and $b = 0.65$ allow a better match to the continuum. Following this procedure, we primarily used the smoothed B-band image to estimate the continuum and we included the V-band to achieve deeper sensitivity and to correct those objects not detected in the B-band image.

Data reduction and analysis for the companion quasar. Upon analysing the continuum-subtracted Lyman-α image, we identified a compact Lyman-α excess source at 24.3 arcsec separation from UM 287 (corresponding to about 200 physical kpc), which has a faint counterpart in our LRIS continuum image and is also detected in the SDSS ($g = 22.8 \pm 0.1$). Further exploration of this source reveals it is detected by the FIRST survey (FIRST J005203.26+010108.6) with a flux $F_{\rm peak} = 21.38$ mJy, strongly suggesting that this source is a radio-loud but optically faint quasar. On UT 08 December 2013, we obtained a long-slit spectrum of J005203.26+010108.6 using the Keck/LRIS spectrometer configured with the D560 dichroic, the 600/4000 grism in the LRISb camera, and the 600/10000 grating in the LRISr camera. We oriented the long slit to also cover UM 287.

These data were reduced with the LowRedux (http://www.ucolick.org/~xavier/LowRedux/index.html) software package using standard techniques. Extended Data Fig. 2 presents the two, optimally extracted spectra from the LRISb camera. One recognizes the broad and bright emission lines characteristic of type I quasars. The redshift estimated from these lines—that has an error of about 800 km s^{-1} (1σ)—is consistent with the systemic redshift of UM 287, suggesting that UM 287 is actually a member of a binary system with a fainter companion. We emphasize, however, that there is very little (if any) Lyman-α emission apparent in the narrow-band image that may be associated with J005203.26+010108.6 apart from that produced by its own nuclear activity.

Because of the large distance from UM 287—at least 200 physical kpc and up to 4 physical Mpc considering the 1σ redshift error—and the morphology of the nebula we can exclude the possibility that the UM 287 nebula is the result of tidal interaction due to a merging event between the two quasar hosts. Indeed, such a large separation would imply that any possible encounter between the two quasars is probably a high velocity interaction or an encounter with large impact parameter. We note that it is not impossible but extremely difficult to produce a long and massive tidal tail during a 'fast' encounter[41], and the amount of gas stripped by the quasar host galaxies in the best scenario would probably be a very small fraction (<10%) of its total interstellar medium. Irrespective of the details of the possible interaction between the two quasar host galaxies, any resulting, long tidal tail would be very thin with sizes of the order of few kpc or less[41] whereas the observed nebula has a FWHM thickness of at least 100 physical kpc in its widest point.

Galaxy overdensity analysis. We have obtained a sample of 60 Lyman-α emitter (LAE) candidates above a flux limit of $3 \times 10^{-18}\,{\rm erg\,s^{-1}\,cm^{-2}}$ (corresponding to a Lyman-α luminosity of about $2 \times 10^{41}\,{\rm erg\,s^{-1}}$) within the volume probed by our narrow-band imaging (~3,100 co-moving Mpc^{-3}) around UM 287. The selection is based on the same technique applied to our pilot survey[5].

How does the number density in our survey compare to other similar searches around massive objects? Surveys of LAE around HzRGs[15,42] have revealed large overdensities of LAEs with respect to field studies at similar redshifts[43,44] that are compatible with the presence of a massive halo as estimated from clustering, that is, $10^{13} M_\odot$. Narrowband imaging of the radio-galaxy MCR 1138-262 at $z = 2.16$ (ref. 42), associated with a 200-kpc-scale Lyman-α nebula, found a number density of LAE above $L_{\rm Ly\alpha} = 1.4 \times 10^{42}\,{\rm erg\,s^{-1}}$ of $n_{\rm HzRG} \approx 10 \pm 2 \times 10^{-3}$ co-moving Mpc^{-3} [15]. By comparison, the number density of LAE above the same limit at the same redshift in the field is $n_{\rm field} \approx (1.5 \pm 0.5) \times 10^{-3}$ co-moving Mpc^{-3}, corrected for completeness[43]. If we restricted our sample to the same luminosity cut, we found a number density of $n_{\rm UM287} \approx (5 \pm 1) \times 10^{-3}$ co-moving Mpc^{-3}. Note that, at this luminosity, our sample is complete. Despite the large statistical errors, we note that the overdensity with respect to the field around UM 287 (about a factor of three) is significantly smaller than the overdensity of LAE around MCR 1138-262 (about a factor six). A similar result is obtained comparing the overdensity of LAE around UM 287 with other HzRGs[15], suggesting that UM 287 is hosted by a smaller halo than typical HzRG hosts. Moreover, the modest overdensity of our field is strong evidence against the possibility that the UM 287 nebula may be fully contained by an individual dark-matter halo of mass $10^{13.5}\,M_\odot$, as would be required by its size. Note that the galaxy number density estimate around

UM 287 is a conservative upper limit: if the quasar is illuminating the surrounding volume, we expect a boost in the number of detectable LAE objects due to fluorescence, as demonstrated in our pilot survey[5]. Our measurement is also compatible with the number density of LAEs found by other recent, shallower surveys for Lyman-α emission around eight radio-quiet, bright quasars[45] at $z \approx 2.7$ that have a host halo mass of $10^{12.5}\ M_\odot$ as constrained by the clustering of Lyman break galaxies. These studies found number densities ranging from 6×10^{-3} to 22×10^{-3} co-moving Mpc^{-3} around individual quasars above a Lyman-α luminosity of $L_{\mathrm{Ly}\alpha} = 5.8 \times 10^{41}\ \mathrm{erg\,s}^{-1}$. Combining the 8 fields, the average number density from their survey is $(12.0 \pm 0.4) \times 10^{-3}$ co-moving Mpc^{-3}.

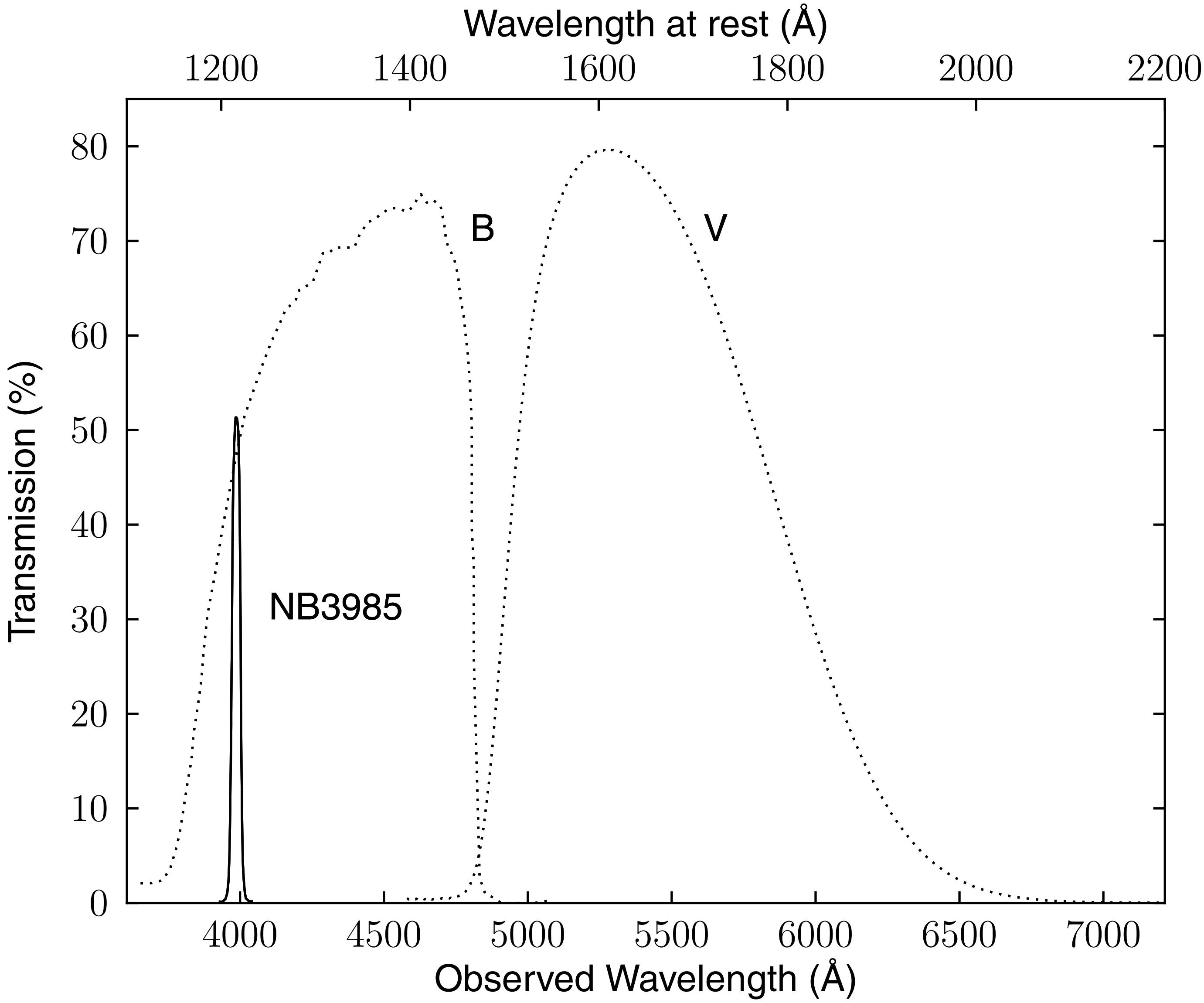
Using the same luminosity cut, we find a number density of $(12 \pm 2) \times 10^{-3}$ co-moving Mpc^{-3}, suggesting that the halo mass of UM 287 is indeed within the typical range for the host haloes of radio-quiet quasars.

Converting the observed Lyman-α emission to gas column densities. We derived a relation between the observable Lyman-α emission from diffuse gas illuminated by a quasar and the gas column densities by combining a Lyman-α radiative transfer model with the results of a cosmological hydrodynamical simulation of structure formation at $z = 2.3$ (ref. 5). The cosmological simulations have been obtained with the adaptive mesh refinement code RAMSES[46] and consist of a 40^3 co-moving Mpc^3 cosmological volume with a 10^3 co-moving Mpc^3 high-resolution region containing a massive halo compatible with the expected quasar hosts, that is, with a dark-matter mass $M_{\mathrm{DM}} \approx 10^{12.5} M_\odot$. The equivalent base-grid resolution in the high-resolution region corresponds to a (1024^3) grid with a dark-matter particle mass of about $1.8 \times 10^6\ M_\odot$. We used other additional 6 grid refinement levels, reaching a maximum spatial resolution of about 0.6 co-moving kpc, that is, about 165 physical pc at $z = 2.3$. Star formation, supernova feedback, and an optically thin ultraviolet background with an on-the-fly self-shielding correction are included using a typical choice of sub-grid parameters for the simulation resolution[5]. We have then applied in post processing an ionization and Lyman-α radiative transfer using the RADAMESH adaptive mesh refinement code[29]. Ionization, Lyman-α and non-ionizing continuum radiation from the quasar broad line region is propagated within two symmetric cones that cover half of the solid angle around the quasar. We included light-travel and finite light-speed effects for both ionizing and Lyman-α radiation transfer and varied the quasar age (from 1 Myr to 10 Myr) and the orientation of the emission cones with respect to the observer line-of-sight and the cosmic web surrounding the simulated halo. We note that these effects are able to produce asymmetric Lyman-α nebulae with sizes and morphologies similar to the observations for short quasar ages (<5 Myr).

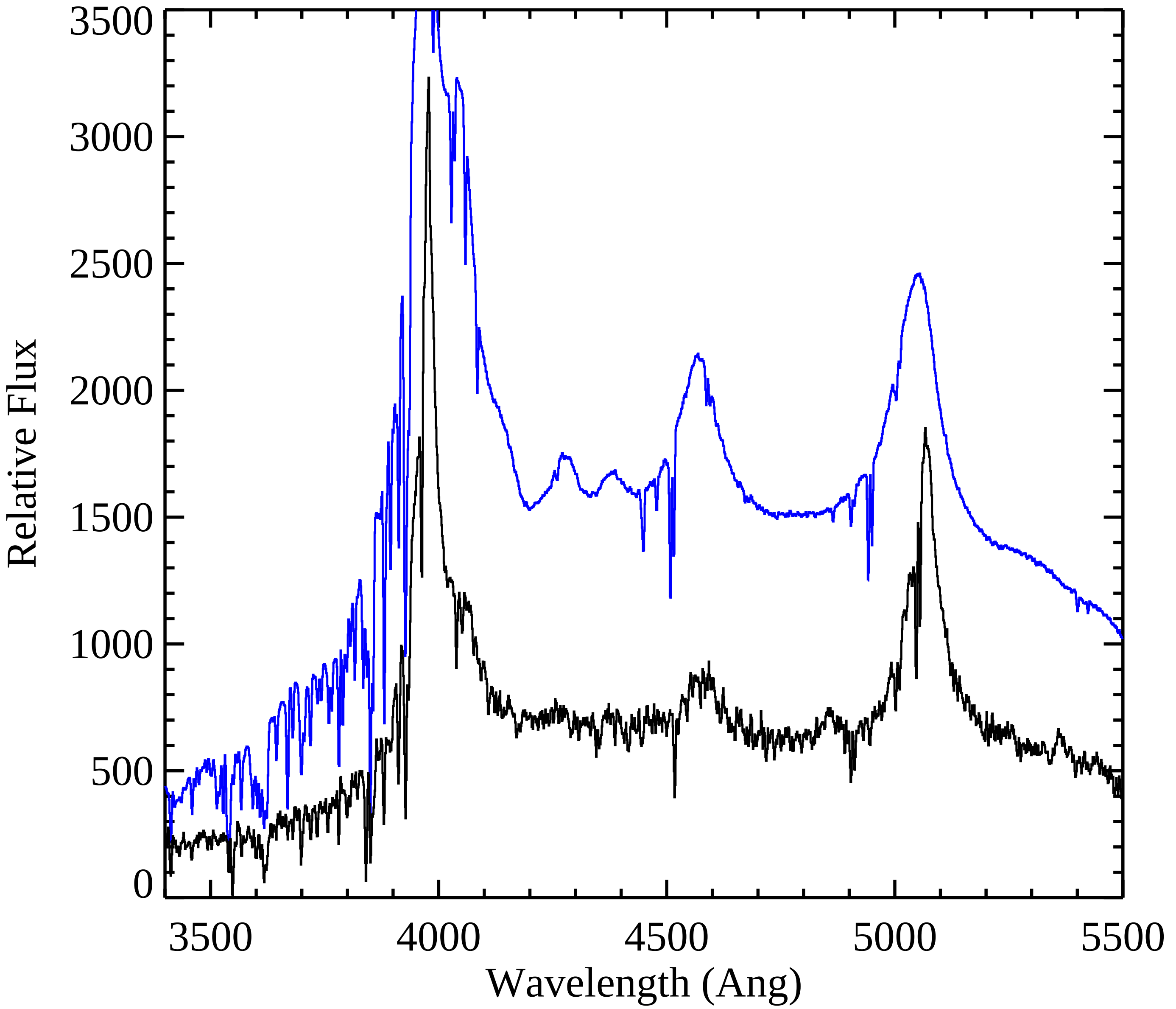
In order to produce a calibrated relation for scenario 1 as discussed in the main text, we have fixed the quasar ionizing and Lyman-α luminosity to the observed value and assumed that the ionizing and Lyman-α emitting cones are coincident. We have then produced mock images with the same angular resolution of the observation that have been convolved with a point spread function (PSF) with 1 arcsec size to simulate atmospheric seeing. A column density map of cold ($T < 5 \times 10^4$ K) ionized hydrogen was produced from the simulations considering only the gas 'illuminated' by the quasar and convolved with the same PSF. We have then cross-correlated the two quantities pixel by pixel and fitted the calibrated relation shown as a solid line in the left panel of Extended Data Fig. 3. This relation is consistent with analytical expectations from highly ionized gas where the Lyman-α emission is mostly produced by hydrogen recombination with a negligible contribution from collisional excitations and Lyman-α scattering (or photon-pumping) from the quasar non-ionizing continuum and Lyman-α radiation[5]. We have repeated the experiment varying the sub-grid clumping factor (C) below the simulation resolution and found, as expected for highly ionized gas, that the simulated surface brightness scales linearly with C at a given gas column density.

We have also considered the extreme case in which the simulated gas is only illuminated by non-ionizing radiation from the quasar, and therefore that dense gas in the simulation remains mostly neutral (scenario 2 in the main text) above the self-shielding density to the cosmic ultraviolet background (about 0.01 atoms cm^{-3}). We obtained and post-processed a mock image as in the previous case and cross-correlated the resulting Lyman-α surface brightness with the neutral hydrogen column densities (N_{HI}). Despite the large scatter, we found a good correlation between these two quantities (right panel of Extended Data Fig. 3) if the surface brightness is normalized by the impact parameter (b) squared. The relation between the Lyman-α surface brightness, neutral hydrogen column density and impact parameter is consistent with simple analytical expectations from pure Lyman scattering from the broad line region of the quasar for Lyman-α optical depth much larger than unity. In this case, the amount of photon-pumping (or, analogously, the equivalent width of the absorbed quasar Lyman-α and continuum emission) is dominated by the line damping wing and therefore is proportional to ${N_{\mathrm{HI}}}^{1/2}$.

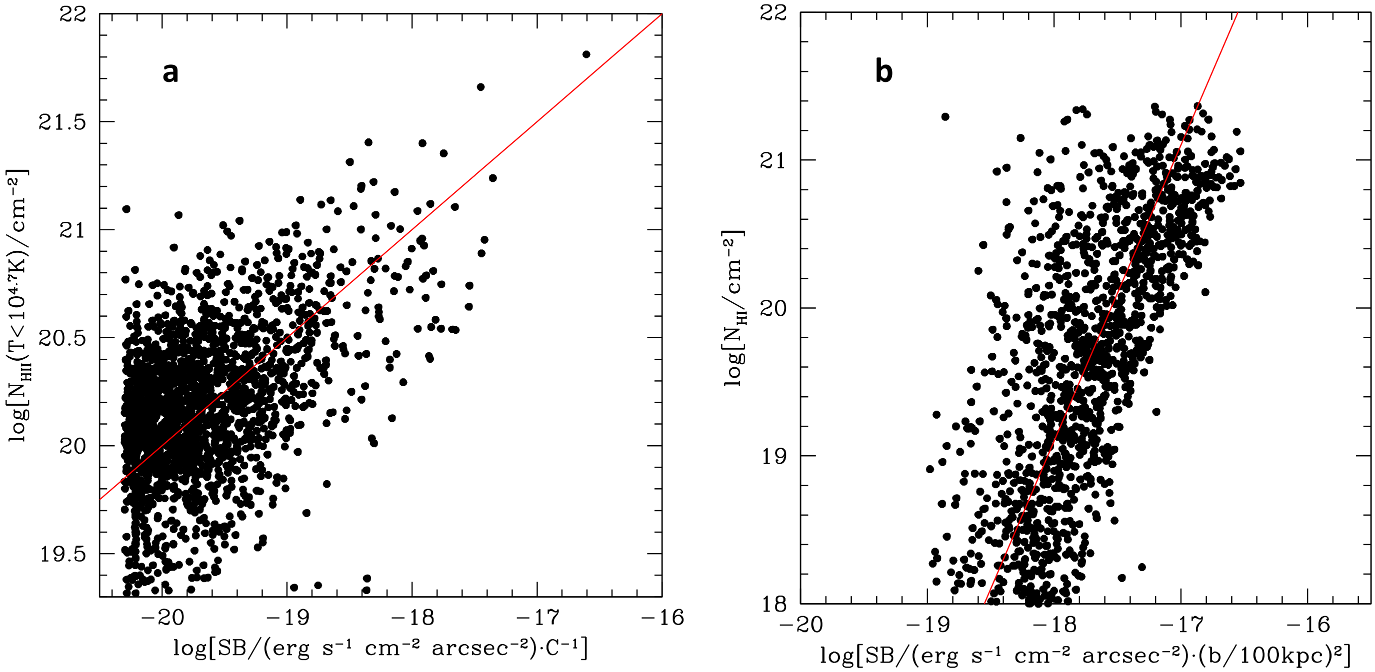
31. McIntosh, D. H., Rieke, M. J., Rix, H.-W., Foltz, C. B. & Weymann, R. J. A statistical study of rest-frame optical emission properties in luminous quasars at $2.0 < z < 2.5$. *Astrophys. J.* **514,** 40–67 (1999).
32. Planck Collaboration et al. Planck 2013 results. XVI. Cosmological parameters. Preprint at http://arxiv.org/abs/1303.5076 (2013).
33. Telfer, R. C., Zheng, W., Kriss, G. A. & Davidsen, A. F. The rest-frame extreme-ultraviolet spectral properties of quasi-stellar objects. *Astrophys. J.* **565,** 773–785 (2002).
34. Hennawi, J. F. *et al.* Quasars probing quasars. I. Optically thick absorbers near luminous quasars. *Astrophys. J.* **651,** 61–83 (2006).
35. Becker, R. H., White, R. L. & Helfand, D. J. in *Astronomical Data Analysis Software and Systems III* (eds Crabtree, D. R., Hanisch, R. J. & Barnes, J.) 165–174 (Astron. Soc. Pacif. Conf. Ser. Vol. 61, 1994).
36. Ivezić, Ž. *et al.* Optical and radio properties of extragalactic sources observed by the FIRST Survey and the Sloan Digital Sky Survey. *Astron. J.* **124,** 2364–2400 (2002).
37. Bertin, E. & Arnouts, S. SExtractor: software for source extraction. *Astron. Astrophys.* **117** (Suppl.), 393–404 (1996).
38. Bertin, E. in *Astronomical Data Analysis Software and Systems XV* (eds Gabriel, C., Arviset, C., Ponz, D. & Enrique, S.) 112–115 (Astron. Soc. Pacif. Conf. Ser. Vol. 351, 2006).
39. Bertin, E. *et al.* in *Astronomical Data Analysis Software and Systems XI* (eds Bohlender, D. A., Durand, D. & Handley, T. H.) 228–237 (Astron. Soc. Pacif. Conf. Ser. Vol. 281, 2002).
40. Landolt, A. U. UBVRI photometric standard stars in the magnitude range 11.5-16.0 around the celestial equator. *Astron. J.* **104,** 340–371 (1992).
41. Barnes, J. E. & Hernquist, L. Dynamics of interacting galaxies. *Annu. Rev. Astron. Astrophys.* **30,** 705–742 (1992).
42. Kurk, J. D. *et al.* A Search for clusters at high redshift. I. Candidate Lyα emitters near 1138-262 at $z = 2.2$. *Astron. Astrophys.* **358,** L1–L4 (2000).
43. Guaita, L. *et al.* Lyα-emitting galaxies at $z = 2.1$ in ECDF-S: building blocks of typical present-day galaxies? *Astrophys. J.* **714,** 255–269 (2010).
44. Ciardullo, R. *et al.* The evolution of Lyα-emitting galaxies between $z = 2.1$ and $z = 3.1$. *Astrophys. J.* **744,** 110 (2012).
45. Trainor, R. F. & Steidel, C. C. Constraints on hyperluminous QSO lifetimes via fluorescent Lyα emitters at $z \sim 2.7$. *Astrophys. J.* **775,** L3 (2013).
46. Teyssier, R. Cosmological hydrodynamics with adaptive mesh refinement. A new high resolution code called RAMSES. *Astron. Astrophys.* **385,** 337–364 (2002).



Extended Data Figure 1 | Measured transmission curves of the filters used in this study. Solid line, NB3985; dotted lines, B band (left) and V band (right). Bottom axis, observed wavelength; top axis, the rest-frame wavelength for sources at $z = 2.27$.



Extended Data Figure 2 | Keck/LRIS spectrum of UM 287 and of the faint, radio-loud companion quasar. Black line, spectrum of this companion quasar which is indicated by 'b' in Fig. 2 and is separated by about 24 arcsec from UM 287. Blue line, spectrum of UM287. Comparison of the two spectra clearly shows that this companion is a quasar at a redshift similar to that of UM 287.



Extended Data Figure 3 | Pixel-to-pixel correlations for Lyman-α surface brightness for scenarios 1 and 2 in the main text. **a**, Pixel-to-pixel correlation between simulated Lyman-α surface brightness (SB) divided by the clumping factor (C) and corresponding cold ($T < 5 \times 10^4$ K) ionized hydrogen column densities N_{HII} for scenario 1 (see text for details). The solid line indicates the relation $N_{\mathrm{HII}} = 10^{21} \times (\mathrm{SB})^{1/2} \times C^{-1/2}$ (here SB is in units of 10^{-18} erg s^{-1} cm^{-2} arcsec^{-2} and C is dimensionless). **b**, Pixel-to-pixel correlation between simulated Lyman-α surface brightness (normalized by the quasar impact parameter squared, b^2) and corresponding neutral hydrogen column density for scenario 2 (see text for details). The solid line represents the relation $N_{\mathrm{HI}} = 10^{19.1} \times [(\mathrm{SB}) \times (b/100)^2]^2$ cm^{-2} (here b is in units of kpc).

LETTER

doi:10.1038/nature12964

Measurement of parity violation in electron–quark scattering

The Jefferson Lab PVDIS Collaboration*

Symmetry permeates nature and is fundamental to all laws of physics. One example is parity (mirror) symmetry, which implies that flipping left and right does not change the laws of physics. Laws for electromagnetism, gravity and the subatomic strong force respect parity symmetry, but the subatomic weak force does not[1,2]. Historically, parity violation in electron scattering has been important in establishing (and now testing) the standard model of particle physics. One particular set of quantities accessible through measurements of parity-violating electron scattering are the effective weak couplings C_{2q}, sensitive to the quarks' chirality preference when participating in the weak force, which have been measured directly[3,4] only once in the past 40 years. Here we report a measurement of the parity-violating asymmetry in electron–quark scattering, which yields a determination of $2C_{2u} - C_{2d}$ (where u and d denote up and down quarks, respectively) with a precision increased by a factor of five relative to the earlier result. These results provide evidence with greater than 95 per cent confidence that the C_{2q} couplings are non-zero, as predicted by the electroweak theory. They lead to constraints on new parity-violating interactions beyond the standard model, particularly those due to quark chirality. Whereas contemporary particle physics research is focused on high-energy colliders such as the Large Hadron Collider, our results provide specific chirality information on electroweak theory that is difficult to obtain at high energies. Our measurement is relatively free of ambiguity in its interpretation, and opens the door to even more precise measurements in the future.

In parity-violating electron scattering (PVES) experiments, an asymmetry is measured that can be expressed as

$$A_{\mathrm{PV}} = \frac{\sigma_{\mathrm{R}} - \sigma_{\mathrm{L}}}{\sigma_{\mathrm{R}} + \sigma_{\mathrm{L}}} \quad (1)$$

where $\sigma_{\mathrm{R}}(\sigma_{\mathrm{L}})$ are the cross-sections for scattering longitudinally polarized electrons that are in the right-handed R (left-handed L) helicity state, meaning their spins are parallel (antiparallel) to the electron's momentum. For deep inelastic scattering (DIS) from nuclear targets (DIS is defined as scattering in which the electron interacts with a single quark, almost independent of the surrounding quarks and gluons), this asymmetry can be written in a largely model-independent way as[5]

$$A_{\mathrm{PV}} = \frac{G_{\mathrm{F}} Q^2}{4\sqrt{2}\pi\alpha}\left[a_1\left(x,Q^2\right)Y_1\left(x,y,Q^2\right) + a_3\left(x,Q^2\right)Y_3\left(x,y,Q^2\right)\right] \quad (2)$$

where G_{F} is the Fermi constant, α is the fine-structure constant, $Q^2 \equiv -q^2$ with q the four-momentum transferred from the electron to the target, x is the Bjorken scaling variable and describes the fraction of momentum carried by the quark struck by the electron, $y = (E - E')/E$ is the fractional energy loss of the electron with $E(E')$ the incident (scattered) electron energy, $Y_{1,3}$ are kinematic factors, and the variables $a_{1,3}$ are related to the subatomic structure of the target. (See Supplementary Methods for a complete description.) The first experiment (SLAC E122) to detect parity violation in electron scattering[3,4] provided results that strongly favoured the model of refs 6–8, establishing it as the keystone of the now highly successful standard model of particle physics. PVES has subsequently been used as a sensitive probe to study diverse physics, ranging from physics beyond the standard model[9,10] to the structure of both nuclei[11] and the nucleon (ref. 12 and references therein).

In so-called tree-level scattering, where the electron exchanges only a single photon or a single Z boson with the target, very simple expressions for $a_{1,3}$ in equation (2) emerge for electron DIS from deuterium:

$$a_1 = \frac{6}{5}\left(2C_{1u} - C_{1d}\right),\ a_3 = \frac{6}{5}\left(2C_{2u} - C_{2d}\right) \quad (3)$$

The use of the deuterium target simplifies the interpretation because it has equal numbers of up and down valence quarks. Here, $C_{1u(1d)}$ and $C_{2u(2d)}$ are the effective weak couplings between the electrons and the up (down) quarks, often collectively written as C_{1q} and C_{2q}. The subscripts 1 and 2 refer to whether the coupling to the electron or quark is vector or axial-vector in nature: $C_{1u(d)}$ is the (AV) combination of the electron's axial-vector weak charge and the quark's vector weak charge, that is, it probes parity violation caused by the difference in the Z^0 coupling between left- and right- handed electron chiral states; $C_{2u(d)}$ is the (VA) combination of the electron's vector weak charge and the quark's axial-vector weak charge that is sensitive to parity violation due to the different quark chiral states. In testing the standard model it is important to determine all four $C_{1u,1d,2u,2d}$ as accurately as possible, because new interactions could manifest themselves in either set of couplings. Experimentally, one could extract both $2C_{1u} - C_{1d}$ and $2C_{2u} - C_{2d}$ by measuring asymmetries at different $Y_{1,3}$ values in the DIS regime. However, a precise determination of $2C_{2u} - C_{2d}$ is difficult because of its small value in the standard model (-0.095), as opposed to $2C_{1u} - C_{1d} = -0.719$.

The new measurement reported here was performed using the electron beam at the Thomas Jefferson National Accelerator Facility (referred to here as Jefferson Lab), in Virginia, USA. A 100-μA, nearly 90%-longitudinally-polarized electron beam was incident on a 20-cm-long liquid deuterium target held at a temperature of 22 K. Scattered particles were detected in a pair of magnetic spectrometers that determined the momentum and the direction of the detected particles to high precision[13]. To directly access $C_{2u,2d}$, the kinematics were chosen so that the bulk of the detected electrons emerged from the target after undergoing a DIS interaction. In contrast, all other PVES experiments after SLAC E122 were performed outside the DIS regime, and thus could not provide clean information on C_{2q}.

The size of the asymmetry expected for this measurement is at the level of 10^{-4}. The major challenge comes from the combination of the high electron event rate, and the high pion background typical of DIS measurements. This was overcome by the use of a custom electronic and data acquisition (DAQ) system with built-in pion rejection capability[14]. The DAQ system successfully counted electrons, event-by-event, at rates up to 600 kHz. The relative uncertainty in the measured asymmetries due to pion background was less than 5×10^{-4}, and that due to counting deadtime was less than 0.4%. The leading systematic uncertainty comes from the normalization by the electron beam polarization, which had a relative uncertainty of (1.2–1.8)%. Beam instability was

*Lists of participants and their affiliations appear at the end of the paper.

not a significant issue because of recent advances in the monitoring and feedback control of the beam, a direct outcome of some of the earlier PVES studies[9–12].

The high intensity of the Jefferson Lab beam allowed the completion of the experiment in just under two months. A total of about 170,000 million scattered electrons were counted at two DIS settings. The asymmetry measured at $E = 6.067$ GeV, $\langle x\rangle = 0.241$, $Y_1 = 1.0$, $Y_3 = 0.44$ and $\langle Q^2\rangle = 1.085$ (GeV c^{-1})^2 was

$$A_{\rm exp} = [-91.1 \pm 3.1(\text{stat.}) \pm 3.0(\text{syst.})] \times 10^{-6} \quad (4)$$

where $\langle x\rangle$ and $\langle Q^2\rangle$ are averaged over the spectrometer acceptance, and stat. and syst. indicate statistical and systematic errors, respectively. This result is to be compared with the standard model (SM) expectation of $A_{\rm SM} = -87.7 \times 10^{-6}$, with an uncertainty of 0.7×10^{-6} dominated by the uncertainty in the parton distribution functions (PDFs), parameterizations of how partons (quarks and gluons) that form the nucleon carry the nucleon's energy. To allow an extraction of $C_{1u,1d}$ and $C_{2u,2d}$, it is necessary to express the asymmetry in terms of these couplings. This relation was calculated using the MSTW2008 leading-order PDF parametrization[15]. For the kinematics above, it gives $A_{\rm SM} = (1.156 \times 10^{-4})[(2C_{1u} - C_{1d}) + 0.348(2C_{2u} - C_{2d})]$, where the relative uncertainties of the coefficients for the $(2C_{1u} - C_{1d})$ and the $(2C_{2u} - C_{2d})$ terms are 0.5% and 5%, respectively. The second DIS setting was at $E = 6.067$ GeV, $\langle x\rangle = 0.295$, $Y_1 = 1.0$, $Y_3 = 0.69$, $\langle Q^2\rangle = 1.901$ (GeV c^{-1})^2, and the result was:

$$A_{\rm exp} = [-160.8 \pm 6.4(\text{stat.}) \pm 3.1(\text{syst.})] \times 10^{-6} \quad (5)$$

The standard model expectation is $A_{\rm SM} = (-158.9 \pm 1.0) \times 10^{-6}$. The coupling sensitivity is $A_{\rm SM} = (2.022 \times 10^{-4})[(2C_{1u} - C_{1d}) + 0.594(2C_{2u} - C_{2d})]$, with the same relative uncertainties as the first DIS setting. Details of the standard model calculation and the uncertainty due to PDF fits are given in Supplementary Methods.

Using the most recent world data for the coupling $C_{1u,1d}$ (ref. 16), obtained from PVES and caesium atomic parity violation experiments[17–20], a simultaneous fit of $2C_{1u} - C_{1d}$ and $2C_{2u} - C_{2d}$ to our results and to the asymmetries from SLAC E122 was performed, yielding:

$$\begin{aligned}(2C_{2u} - C_{2d})|_{Q^2=0} &= -0.145 \pm 0.066(\text{exp.}) \\ &\quad \pm 0.011(\text{PDF}) \pm 0.012(\text{HT}) \\ &= -0.145 \pm 0.068(\text{total})\end{aligned} \quad (6)$$

Here, exp. refers to the total experimental uncertainty, given by the statistical and the systematic uncertainties of the asymmetry results added in quadrature. The third uncertainty is due to the so-called higher-twist (HT) effects, caused by interactions among quarks inside the target. Further theoretical uncertainties, including QED vacuum polarization and the γZ box diagram, are negligible compared to the uncertainty due to the PDF fits. Electroweak and process-specific radiative corrections have been applied to calculate the values at zero-Q^2, $C_{2u,2d}|_{Q^2=0}$ called $g_{\rm VA}^{eu,ed}$ with e referring to electrons (and similarly $C_{1u,1d}|_{Q^2=0}$ called $g_{\rm AV}^{eu,ed}$) in ref. 21, so that the values in equation (6) can be compared directly to results from other precision experiments using different kinds of processes. The values for $C_{2u,2d}|_{Q^2=0}$ differ from those at both Q^2 accessed in this experiment by 0.002–0.003 for both the up and the down quarks.

The asymmetry results in equations (4) and (5) can also be interpreted as a determination of the weak mixing angle θ_W, an important ingredient of the electroweak unification of the standard model. The result, evolved to the mass of the Z boson in the modified minimal subtraction ($\overline{\rm MS}$) scheme, is $\hat{s}_Z^2 \equiv \sin^2\theta_{W(Q^2=M_Z^2,\overline{\rm MS})} = 0.2299 \pm 0.0043$, in agreement with the latest standard-model fit to world data, $\hat{s}_Z^2 = 0.23126 \pm 0.00005$.

The result in equation (6) is compared with the standard-model prediction $2C_{2u} - C_{2d}|_{Q^2=0} = -0.0950 \pm 0.0004$ in Fig. 1. Our results have greatly improved the uncertainty on the effective electron–quark VA weak couplings $C_{2u,2d}$ and are in good agreement with the standard-model prediction. This is also the first direct measurement of the coupling combination $2C_{2u} - C_{2d}$ that deviates from zero. We note that evidence for non-zero values of the $C_{2u,2d}$, possibly in a different combination from what we measured, may have been observed in experiments measuring the nucleon axial form factors[22]. However, extraction of $C_{2u,2d}$ from the nucleon axial form factor is model-dependent, whereas in DIS the electron probes quarks unambiguously. The directness of our approach is essential to reach a significantly higher accuracy in the future, such as through the PVDIS programme planned for the 12 GeV upgrade of Jefferson Lab.

A comparison of the present result with the standard-model predictions can be used to set mass limits Λ below which new interactions are unlikely to occur. For the cases of electron and quark compositeness and contact interactions, we used the convention of ref. 23 and the procedure in ref. 24. The limit for the constructive (destructive) interference contribution to the standard model is:

$$\Lambda^{\pm} = {\rm v}\left[\frac{8\sqrt{5}\pi}{\left|(2C_{2u} - C_{2d})_{Q^2=0}\right|^{\pm}}\right]^{1/2} \quad (7)$$

where $\left|(2C_{2u} - C_{2d})_{Q^2=0}\right|^{\pm}$ is the difference between the standard-model value and the upper (lower) confidence bound of the data, $v = \sqrt{\sqrt{2}/(2G_{\rm F})} = 246.22$ GeV is the Higgs vacuum expectation value setting the electroweak scale, and the $\sqrt{5}$ is a normalization factor taking into account the coefficients of the $C_{2u,2d}$ in the denominator.

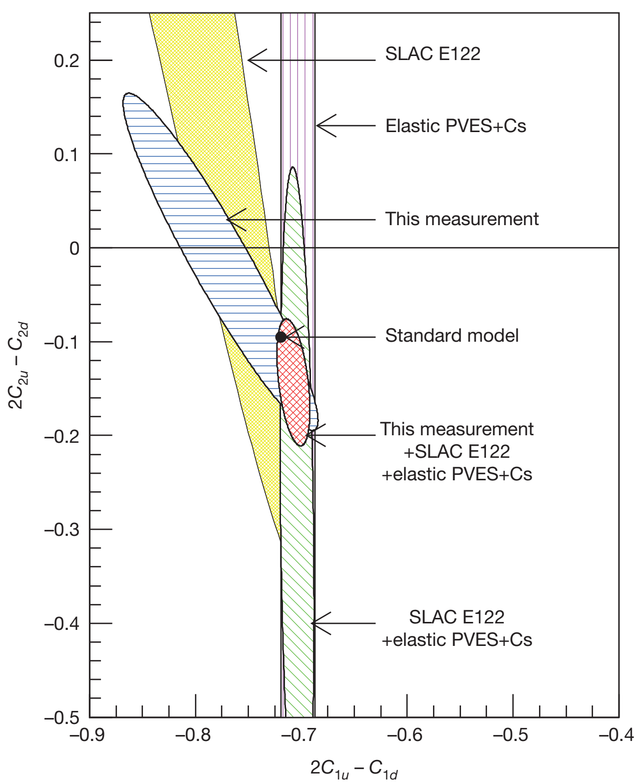


Figure 1 | Comparison of the present results with those of earlier experiments and predictions of the standard model. Values of $(2C_{1u} - C_{1d})|_{Q^2=0}$ and $(2C_{2u} - C_{2d})|_{Q^2=0}$ from this experiment (ellipse with blue horizontal hatching) are compared with those of SLAC E122 (yellow ellipse)[3,4]. The latest data on C_{1q} (from PVES[16] and atomic Cs[17–20]) are shown as the band with magenta vertical hatching. The ellipse with diagonal green hatching shows the combined result of SLAC E122 and the latest C_{1q}, while the ellipse with red cross-hatching shows the combined result of SLAC E122, this experiment, and the latest C_{1q}. The standard model value (with negligible uncertainty) is shown as the black dot, where the size of the dot is for visibility.

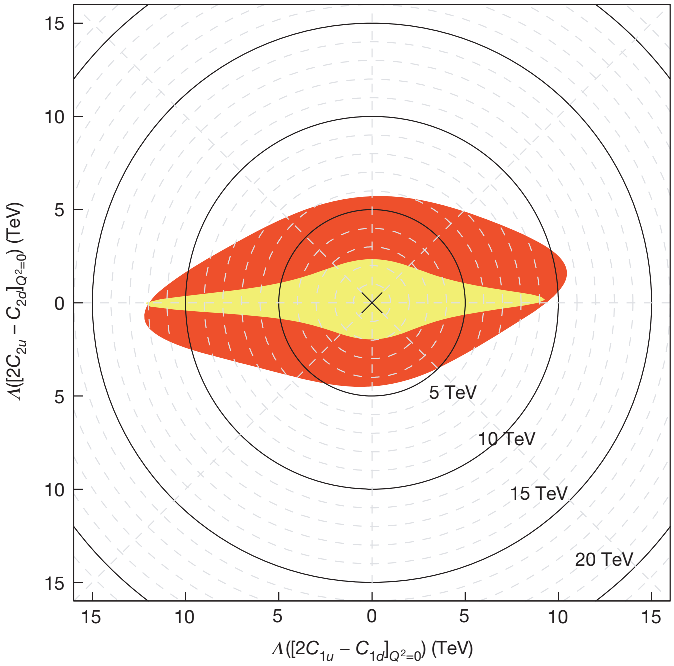


Figure 2 | Mass exclusion limits Λ on the electron and quark compositeness and contact interactions. These limits are obtained from the zero-Q^2 values of $2C_{1u} - C_{1d}$ and $2C_{2u} - C_{2d}$ at the 95% confidence level. The outside of the yellow shape shows the limit obtained from SLAC E122 asymmetry results[3,4] combined with the best C_{1q} values[16]. The outside of the red shape shows the limit with our new results added. For visual guidance, mass limit scales in TeV are shown as solid and dashed circles.

For a 95% confidence level, we extracted

$$\Lambda^+ = 5.8 \text{ TeV} \quad \text{and} \quad \Lambda^- = 4.6 \text{ TeV} \tag{8}$$

for constructive and destructive interference from beyond-the-standard-model physics. Figure 2 illustrates these limits. The limits set by $C_{1u,1d}$ are determined mostly by previous PVES and caesium atomic-parity-violation results, but this experiment clearly improves the limits set by $C_{2u,2d}$.

The strength of our results reported here is that they isolate a well-defined combination of the electron–quark contact interactions. We note that mass limits on the electron–quark contact interactions have been published by the ZEUS[25] and H1[26] collaborations at the Hadron–Electron Ring Accelerator, HERA. They find $\Lambda^+ = 3.3$ TeV and $\Lambda^- = 3.2$ TeV (ref. 25), and $\Lambda^+ = 3.8$ TeV and $\Lambda^- = 3.6$ TeV (ref. 26), respectively, on the electron–quark VA term. Similar limits of $\Lambda^+ = 9.5$ TeV and $\Lambda^- = 12.1$ TeV have been published by the ATLAS collaboration[27] at the LHC in the left-left isoscalar model. To account for the different chirality structure of the models used, the HERA limits on the electron–quark VA model need to be scaled by $2^{-1/4} = 0.84$, while the LHC limits using the left-left isoscalar model need to be scaled by $2^{1/4} = 1.19$, in order to be compared to the mass limits extracted from $C_{2u,2d}$. The HERA and the LHC measurements are sensitive to several different vector and axial-vector weak charge combinations, thus their limits were obtained with the assumption that, apart from the particular chirality combination used in the model, all other contact interactions are zero. This assumption is unnecessary for the extraction of mass limits from our results. The chiral structure of the effective electron–quark weak couplings C_{2q} isolates interactions beyond the standard model in which it is the chirality of the quarks that is responsible for the observed parity violation.

METHODS SUMMARY

The parity-violating asymmetry A_{exp} between right- and left-handed electrons was computed from the detected counts C, normalized by the beam intensity I, and integrated over periods of stable beam helicity. Two kinds of corrections were then made to the asymmetries: overall normalization factors and possible systematic shifts due to false asymmetries arising from backgrounds or helicity correlations in the beam parameters. The normalization factors include the beam polarization, measurements of scattered-electron kinematics, electromagnetic radiative corrections, and effects from two-photon exchange between the electron and target. The false-asymmetry corrections were all very small compared to the statistical error and included an evaluation of helicity correlations in beam current, position and energy, and backgrounds such as pions, scattering from the target aluminium windows, or rescattering inside the spectrometers. A summary of all corrections and the asymmetry results is presented in Supplementary Table 1.

To calculate the standard-model expectation of the measured asymmetry and its sensitivity to $2C_{1u} - C_{1d}$ and $2C_{2u} - C_{2d}$, we used PDFs to calculate the structure functions in $a_{1,3}$. Three PDF fits were used. Results of the calculation are shown in Supplementary Table 2. The relative variation among all three fits is less than 0.6% for the a_1 term, and less than 5% for the a_3 term of the asymmetry. Effects from interactions among quarks inside the target, called 'higher-twist effects', were evaluated using the most recent theoretical bounds combined with data on neutrino structure functions. It was found that the uncertainty in the extraction of $2C_{2u} - C_{2d}$ due to the higher-twist effects is at the same level as that due to the PDFs, and is quite small compared to the experimental uncertainties.

Received 28 October; accepted 17 December 2013.

1. Lee, T. D. & Yang, C.-N. Question of parity conservation in weak interactions. *Phys. Rev.* **104,** 254–258 (1956).
2. Wu, C. S., Ambler, E., Hayward, R. W., Hoppes, D. D. & Hudson, R. P. Experimental test of parity conservation in beta decay. *Phys. Rev.* **105,** 1413–1415 (1957).
3. Prescott, C. Y. *et al.* Parity nonconservation in inelastic electron scattering. *Phys. Lett. B* **77,** 347–352 (1978).
4. Prescott, C. Y. *et al.* Further measurements of parity nonconservation in inelastic electron scattering. *Phys. Lett. B* **84,** 524–528 (1979).
5. Cahn, R. N. & Gilman, F. J. Polarized-electron-nucleon scattering in gauge theories of weak and electromagnetic interactions. *Phys. Rev. D* **17,** 1313–1322 (1978).
6. Glashow, S. L. Partial symmetries of weak interactions. *Nucl. Phys.* **22,** 579–588 (1961).
7. Weinberg, S. A model of leptons. *Phys. Rev. Lett.* **19,** 1264–1266 (1967).
8. Salam, A. Weak and electromagnetic interactions. In *Elementary Particle Theory: Relativistic Groups and Analyticity* (ed. Svartholm, N.) 367–377 (Almqvist and Wiksell, 1968).
9. Anthony, P. L. *et al.* Precision measurement of the weak mixing angle in Moller scattering. *Phys. Rev. Lett.* **95,** 081601 (2005).
10. Czarnecki, A. & Marciano, W. J. Electrons are not ambidextrous. *Nature* **435,** 437–438 (2005).
11. Abrahamyan, S. *et al.* Measurement of the neutron radius of ^{208}Pb through parity-violation in electron scattering. *Phys. Rev. Lett.* **108,** 112502 (2012).
12. Armstrong, D. S. & McKeown, R. D. Parity-violating electron scattering and the electric and magnetic strange form factors of the nucleon. *Annu. Rev. Nucl. Part. Sci.* **62,** 337–359 (2012).
13. Alcorn, J. *et al.* Basic instrumentation for Hall A at Jefferson Lab. *Nucl. Instrum. Methods A* **522,** 294–346 (2004).
14. Subedi, R. *et al.* A scaler-based data acquisition system for measuring parity violation asymmetry in deep inelastic scattering. *Nucl. Instrum. Methods A* **724,** 90–103 (2013).
15. Martin, A. D., Stirling, W. J., Thorne, R. S. & Watt, G. Parton distributions for the LHC. *Eur. Phys. J. C* **63,** 189–285 (2009).
16. Androic, D. *et al.* First determination of the weak charge of the proton. *Phys. Rev. Lett.* **111,** 141803 (2013).
17. Wood, C. S. *et al.* Measurement of parity nonconservation and an anapole moment in cesium. *Science* **275,** 1759–1763 (1997).
18. Bennett, S. C. & Wieman, C. E. Measurement of the $^6S \rightarrow ^7S$ transition polarizability in atomic cesium and an improved test of the Standard Model. *Phys. Rev. Lett.* **82,** 2484–2487 (1999); erratum **83,** 889 (1999).
19. Ginges, J. S. M. & Flambaum, V. V. Violations of fundamental symmetries in atoms and tests of unification theories of elementary particles. *Phys. Rep.* **397,** 63–154 (2004).
20. Dzuba, V. A., Berengut, J. C., Flambaum, V. V. & Roberts, B. Revisiting parity nonconservation in cesium. *Phys. Rev. Lett.* **109,** 203003 (2012).
21. Erler, J. & Su, S. The weak neutral current. *Prog. Part. Nucl. Phys.* **71,** 119–149 (2013).
22. Beise, E. J., Pitt, M. L. & Spayde, D. T. The SAMPLE experiment and weak nucleon structure. *Prog. Part. Nucl. Phys.* **54,** 289–350 (2005).
23. Eichten, E., Lane, K. D. & Peskin, M. E. New tests for quark and lepton substructure. *Phys. Rev. Lett.* **50,** 811–814 (1983).
24. Schael, S. *et al.* Electroweak measurements in electron-positron collisions at *W*-boson-pair energies at LEP. *Phys. Rep.* **532,** 119–244 (2013).
25. Chekanov, S. *et al.* Search for contact interactions, large extra dimensions and finite quark radius in *ep* collisions at HERA. *Phys. Lett. B* **591,** 23–41 (2004).
26. Aaron, F. D. *et al.* Search for contact interactions in $e^{\pm}p$ collisions at HERA. *Phys. Lett. B* **705,** 52–58 (2011).

27. Aad, G. *et al.* Search for contact interactions and large extra dimensions in dilepton events from *pp* collisions at $\sqrt{s}=7$ TeV with the ATLAS detector. *Phys. Rev. D* **87,** 015010 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the personnel of Jefferson Lab for their efforts which resulted in the successful completion of the experiment, and A. Accardi, P. Blunden, W. Melnitchouk and their collaborators for carrying out the calculations necessary for the completion of the data analysis. X.Z. thanks the Medium Energy Physics Group at the Argonne National Laboratory for support during the initial work on this experiment. J.E. was supported by PAPIIT (DGAPAUNAM) project IN106913 and CONACyT (México) project 151234, and acknowledges the hospitality and support by the Mainz Institute for Theoretical Physics (MITP) where part of his work was completed. This work was supported in part by the Jeffress Memorial Trust (award no. J-836), the US NSF (award no. 0653347), and the US DOE (award nos DE-SC0003885 and DE-AC02-06CH11357). This work was authored by Jefferson Science Associates, LLC under US DOE contract no. DE-AC05-06OR23177. The US Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce this manuscript for US Government purposes.

Author Contributions Authors contributed to one or more of the following areas: proposing, leading, and running the experiment; design, construction, optimization, and testing of the data acquisition system; data analysis; simulation; extraction of the physics results from measured asymmetries; and the writing of this Letter.

Author Information J.E. is currently on sabbatical leave at the PRISMA Cluster of Excellence and MITP, Johannes Gutenberg University. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.Z. (xz5y@virginia.edu).

The Jefferson Lab PVDIS Collaboration

D. Wang[1], K. Pan[2], R. Subedi[1]†, X. Deng[1], Z. Ahmed[3], K. Allada[4], K. A. Aniol[5], D. S. Armstrong[6], J. Arrington[7], V. Bellini[8], R. Beminiwattha[9], J. Benesch[10], F. Benmokhtar[11], W. Bertozzi[2], A. Camsonne[10], M. Canan[12], G. D. Cates[1], J.-P. Chen[10], E. Chudakov[10], E. Cisbani[13], M. M. Dalton[1], C. W. de Jager[1,10], R. De Leo[14], W. Deconinck[2]†, A. Deur[10], C. Dutta[4], L. El Fassi[15], J. Erler[16], D. Flay[17], G. B. Franklin[11], M. Friend[11], S. Frullani[13], F. Garibaldi[13], S. Gilad[2], A. Giusa[8], A. Glamazdin[18], S. Golge[12], K. Grimm[19], K. Hafidi[7], J.-O. Hansen[10], D. W. Higinbotham[10], R. Holmes[3], T. Holmstrom[20], R. J. Holt[7], J. Huang[2], C. E. Hyde[12,21], C. M. Jen[3], D. Jones[1], Hoyoung Kang[22], P. M. King[9], S. Kowalski[2], K. S. Kumar[23], J. H. Lee[6,9], J. J. LeRose[10], N. Liyanage[1], E. Long[24], D. McNulty[23]†, D. J. Margaziotis[5], F. Meddi[25], D. G. Meekins[10], L. Mercado[23], Z.-E. Meziani[17], R. Michaels[10], M. Mihovilovic[26], N. Muangma[2], K. E. Myers[27]†, S. Nanda[10], A. Narayan[28], V. Nelyubin[1], Nuruzzaman[28], Y. Oh[22], D. Parno[11], K. D. Paschke[1], S. K. Phillips[29], X. Qian[30], Y. Qiang[30], B. Quinn[11], A. Rakhman[3], P. E. Reimer[7], K. Rider[20], S. Riordan[1], J. Roche[9], J. Rubin[7], G. Russo[8], K. Saenboonruang[1]†, A. Saha[10]‡, B. Sawatzky[10], A. Shahinyan[31], R. Silwal[1], S. Sirca[26], P. A. Souder[3], R. Suleiman[10], V. Sulkosky[2], C. M. Sutera[8], W. A. Tobias[1], G. M. Urciuoli[25], B. Waidyawansa[9], B. Wojtsekhowski[10], L. Ye[32], B. Zhao[6] & X. Zheng[1]

[1]University of Virginia, Charlottesville, Virginia 22904, USA. [2]Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Syracuse University, Syracuse, New York 13244, USA. [4]University of Kentucky, Lexington, Kentucky 40506, USA. [5]California State University, Los Angeles, Los Angeles, California 90032, USA. [6]College of William and Mary, Williamsburg, Virginia 23187, USA. [7]Physics Division, Argonne National Laboratory, Argonne, Illinois 60439, USA. [8]Istituto Nazionale di Fisica Nucleare, Dipt. di Fisica dell'Univ. di Catania, I-95123 Catania, Italy. [9]Ohio University, Athens, Ohio 45701, USA. [10]Thomas Jefferson National Accelerator Facility, Newport News, Virginia 23606, USA. [11]Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. [12]Old Dominion University, Norfolk, Virginia 23529, USA. [13]INFN, Sezione di Roma, gruppo Sanità and Istituto Superiore di Sanità, I-00161 Rome, Italy. [14]Università di Bari, I-70126 Bari, Italy. [15]Rutgers, The State University of New Jersey, Newark, New Jersey 07102, USA. [16]Instituto de Física, Universidad Nacional Autónoma de México, 04510 México D.F., Mexico. [17]Temple University, Philadelphia, Pennsylvania 19122, USA. [18]Kharkov Institute of Physics and Technology, Kharkov 61108, Ukraine. [19]Louisiana Technical University, Ruston, Louisiana 71272, USA. [20]Longwood University, Farmville, Virginia 23909, USA. [21]Clermont Université, Université Blaise Pascal, CNRS/IN2P3, Laboratoire de Physique Corpusculaire, FR-63000 Clermont-Ferrand, France. [22]Seoul National University, Seoul 151-742, South Korea. [23]University of Massachusetts Amherst, Amherst, Massachusetts 01003, USA. [24]Kent State University, Kent, Ohio 44242, USA. [25]INFN, Sezione di Roma and Sapienza — Università di Roma, I-00161 Rome, Italy. [26]Institut Jožef Stefan, SI-1001 Ljubljana, Slovenia. [27]George Washington University, Washington DC 20052, USA. [28]Mississippi State University, Starkeville, Mississippi 39762, USA. [29]University of New Hampshire, Durham, New Hampshire 03824, USA. [30]Duke University, Durham, North Carolina 27708, USA. [31]Yerevan Physics Institute, Yerevan 0036, Armenia. [32]China Institute of Atomic Energy, Beijing 102413, China. †Present addresses: Richland College, Dallas County Community College District, Dallas, Texas 75243, USA (R.S.); College of William and Mary, Williamsburg, Virginia 23187, USA (W.D.); Idaho State University, Pocatello, Idaho 83201, USA (D.M.); Rutgers, The State University of New Jersey, Newark, New Jersey 07102, USA (K.E.M.); Kasetsart University, Bangkok 10900, Thailand (K.S.).

‡Deceased.

doi:10.1038/nature12941

An optical lattice clock with accuracy and stability at the 10^{-18} level

B. J. Bloom[1,2]*, T. L. Nicholson[1,2]*, J. R. Williams[1,2]†, S. L. Campbell[1,2], M. Bishof[1,2], X. Zhang[1,2], W. Zhang[1,2], S. L. Bromley[1,2] & J. Ye[1,2]

Progress in atomic, optical and quantum science[1,2] has led to rapid improvements in atomic clocks. At the same time, atomic clock research has helped to advance the frontiers of science, affecting both fundamental and applied research. The ability to control quantum states of individual atoms and photons is central to quantum information science and precision measurement, and optical clocks based on single ions have achieved the lowest systematic uncertainty of any frequency standard[3–5]. Although many-atom lattice clocks have shown advantages in measurement precision over trapped-ion clocks[6,7], their accuracy has remained 16 times worse[8–10]. Here we demonstrate a many-atom system that achieves an accuracy of 6.4×10^{-18}, which is not only better than a single-ion-based clock, but also reduces the required measurement time by two orders of magnitude. By systematically evaluating all known sources of uncertainty, including *in situ* monitoring of the blackbody radiation environment, we improve the accuracy of optical lattice clocks by a factor of 22. This single clock has simultaneously achieved the best known performance in the key characteristics necessary for consideration as a primary standard—stability and accuracy. More stable and accurate atomic clocks will benefit a wide range of fields, such as the realization and distribution of SI units[11], the search for time variation of fundamental constants[12], clock-based geodesy[13] and other precision tests of the fundamental laws of nature. This work also connects to the development of quantum sensors and many-body quantum state engineering[14] (such as spin squeezing) to advance measurement precision beyond the standard quantum limit.

Accuracy for the SI (International System of Units) second is currently defined by the caesium (Cs) primary standard. However, optical atomic clocks have now achieved a lower systematic uncertainty[3–5,8,12]. This systematic uncertainty will become accuracy once the SI second has been redefined. Neutral atom clocks with many ultracold atoms confined in magic-wavelength optical lattices[15] have the potential for much greater precision than ion clocks[7–9,16]. This potential has been realized only very recently owing to the improved frequency stability of optical local oscillators[14,17,18], resulting in a record single-clock instability of $3.1 \times 10^{-16}/\sqrt{\tau}$, where τ is the averaging time in seconds[6]. This result represents a gain by a factor of 10 in our clock stability, allowing for a factor-of-100 reduction in the averaging time that is required to reach a desired uncertainty[6]. Equivalent instability at one second has also been recently achieved with ytterbium (Yb) optical lattice clocks[7] and averaging for seven hours was demonstrated, down to about 2×10^{-18} for a single clock. We used this measurement precision to evaluate the important systematic effects that have limited optical lattice clocks, and we achieve a total systematic uncertainty in fractional frequency of 6.4×10^{-18}, which is a factor-of-22 improvement over the best published total uncertainties for optical lattice clocks[8–10].

Now that the clock systematic uncertainty has been fully evaluated, it is a frequency standard at which the statistical uncertainty matches the total systematic uncertainty within 3,000 s. Combining improved clock designs with this measurement precision has allowed us to overcome two main obstacles to achieve the reductions in uncertainty reported here. First, we must understand and overcome the atomic-interaction-induced frequency shifts inherent in many-particle clocks[19–21]. We have now determined this effect with 6×10^{-19} uncertainty. Second, we need to measure the thermal radiation environment of the lattice-trapped atoms accurately, because this causes the largest systematic clock shift, known as the blackbody radiation (BBR) Stark shift. Incomplete knowledge of the thermal radiation impinging upon the atoms has so far dominated lattice clock uncertainty. We demonstrate that a combination of accurate *in situ* temperature probes and a thermal enclosure surrounding the clock vacuum chamber allows us to achieve an overall BBR shift uncertainty of 4.1×10^{-18}. This progress was enabled by a precise measurement (performed at the Physikalisch-Technische Bundesanstalt) of the Sr polarizability[22], which governs the magnitude of the BBR shift. Furthermore, we compared two independent Sr clocks and they agree within their combined total uncertainty of 5.4×10^{-17} over a period of one month.

To demonstrate the improved performance of lattice clocks, we built two Sr clocks in JILA[6,23] (see the Methods Summary for details). Herein we refer to the first-generation JILA Sr clock as SrI and the newly constructed Sr clock as SrII. The recent improvement of low-thermal-noise optical oscillators allowed us to demonstrate the stability of both Sr clocks, reaching within a factor of 2 of the quantum projection noise limit for 2,000 atoms[6]. We constructed the SrII clock with the goal of reducing the atomic-interaction-related and BBR-related frequency uncertainties. Thus, SrII has an optical trap volume about 100 times larger than that of SrI to reduce the atomic density, along with *in situ* BBR probes in vacuum to measure the thermal environment of the atoms, achieving a total systematic uncertainty of 6.4×10^{-18}. The improvement of SrI, on the other hand, has been a modest factor of 2 over our previous result[8], now achieving a total systematic uncertainty of 5.3×10^{-17}.

A major practical concern is the speed with which these clocks reach agreement at their stated uncertainties. Hence, the low instability of these Sr clocks (3×10^{-18} at about 10,000 s), displayed as the Allan deviation of their frequency comparison in Fig. 1a, is critical for evaluating systematic effects in a robust manner. Figure 1b documents a comparison of the SrI and SrII clocks over a period of one month, showing that their measured disagreement of $\nu_{\mathrm{SrII}} - \nu_{\mathrm{SrI}} = -2.8 \times 10^{-17}$, with 2×10^{-18} statistical uncertainty, is within their combined systematic uncertainty of 5.4×10^{-17}. The Allan deviation and the binned intercomparison data showcase the stability and reproducibility of these clocks on both short and long timescales. This performance level is necessary for a rigorous evaluation of clock systematics at the 10^{-18} level.

SrI and SrII independently correct for systematic offsets to their measured atomic frequencies. Table 1 lists the major sources of frequency shifts Δ and their related uncertainties σ that affect both clocks. The SrI clock uncertainty is dominated by its BBR shift uncertainty of 4.5×10^{-17}.

[1]JILA, National Institute of Standards and Technology and University of Colorado, Boulder, Colorado 80309-0440, USA. [2]Department of Physics, University of Colorado, Boulder, Colorado 80309-0390, USA. †Present address: Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA.
*These authors contributed equally to this work.

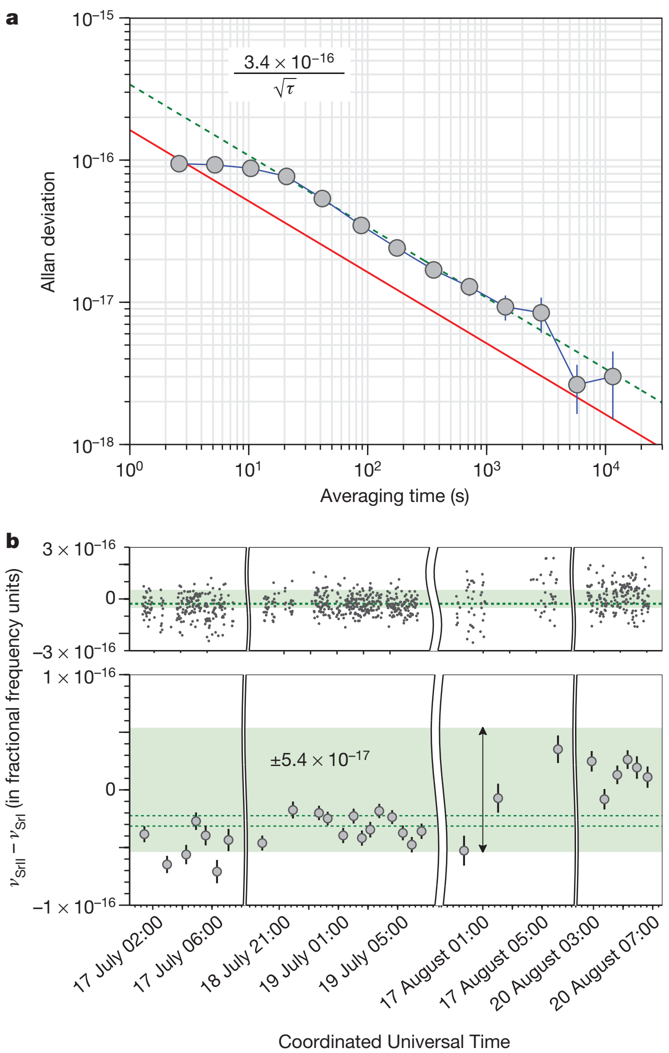


Figure 1 | **Clock comparisons between SrI and SrII. a**, Allan deviation of the SrI and SrII comparison divided by $\sqrt{2}$ to reflect the performance of a single clock. The red solid line is the calculated quantum projection noise for this comparison. The green dashed line is a fit to the data, showing the worst case scenario for the averaging of a single clock of 3.4×10^{-16} at one second. The vertical blue lines represent the 1σ standard errors for the Allan deviation. **b**, The absolute agreement between SrI and SrII recorded at the indicated Coordinated Universal Time. The light-green region denotes the 1σ combined systematic uncertainty for the two clocks under the running conditions at that time. The top panel shows the frequency record binned at 60 s; in the bottom panel each solid circle represents 30 min of averaged data. In the bottom panel, small solid black lines represent the 1σ standard errors inflated by the square root of the reduced chi-squared, $\sqrt{\chi^2_{\text{reduced}}}$. For clarity, we have omitted the error bars in the top panel. The green dashed lines represent the 1σ standard error inflated by the square root of the reduced chi-squared for the weighted mean of these binned comparison data. The final comparison over 52,000 s of data showed agreement at $-2.7(5) \times 10^{-17}$ ($\sqrt{\chi^2_{\text{reduced}}} = 10.5$) for the 30-min averaging time and $-2.8(2) \times 10^{-17}$ ($\sqrt{\chi^2_{\text{reduced}}} = 3.5$) for the 60-s averaging time (see Methods).

For SrII, on the other hand, all sources have been evaluated to produce uncertainties better than 4×10^{-18}.

The largest improvement (compared to other lattice clocks) in the total systematic uncertainty of SrII was obtained through control of the BBR shift. We enclosed the entire clock apparatus inside a BBR shielding box (Fig. 2a). Our lasers for cooling, trapping and clock spectroscopy are delivered to the inside of the BBR shielding box by optical fibres, preventing stray radiation from entering. We have also installed two

Table 1 | **Frequency shifts and related uncertainties for SrI and SrII**

Sources for shift	Δ_{SrI}	σ_{SrI}	Δ_{SrII}	σ_{SrII}
BBR static	−4,832	45	−4,962.9	1.8
BBR dynamic	−332	6	−345.7	3.7
Density shift	−84	12	−4.7	0.6
Lattice Stark	−279	11	−461.5	3.7
Probe beam a.c. Stark	8	4	0.8	1.3
First-order Zeeman	0	<0.1	−0.2	1.1
Second-order Zeeman	−175	1	−144.5	1.2
Residual lattice vector shift	0	<0.1	0	<0.1
Line pulling and tunnelling	0	<0.1	0	<0.1
d.c. Stark	−4	4	−3.5	2.1
Background gas collisions	0	0.07	0	0.6
AOM phase chirp	−7	20	0.6	0.4
Second-order Doppler	0	<0.1	0	<0.1
Servo error	1	4	0.4	0.6
Totals	−5,704	53	−5,921.2	6.4

Shifts and uncertainties are given in fractional frequency units multiplied by 10^{-18}. Uncertainties are quoted as 1σ standard errors. They are determined with the square root of the quadrature sum of the systematic error and statistical error, with the latter quantity inflated by $\sqrt{\chi^2_{\text{reduced}}}$. For SrI, the significant digit for each uncertainty ends at the 1×10^{-18} level; for SrII, the significant digit is extended to the 1×10^{-19} level. See the text and Methods for a detailed discussion of all these systematic uncertainties, including the hyperpolarizability effect of the lattice Stark shift.

in situ silicon diode temperature sensors (with calibrations traceable to the National Institute of Standards and Technology (NIST)) near the atoms to measure their radiative heat environment (Fig. 2a). The sensors were affixed to separate glass tubes (Fig. 2a), which prevented parasitic heat conduction from the chamber to the sensors by providing insulation and radiative dissipation of conductive heat. To improve the radiative coupling, the surfaces of the sensors were coated with high absorptivity, ultrahigh-vacuum-compatible paint. One sensor was mounted 2.54 cm away from the atoms and provided real-time temperature monitoring during clock operation. The second sensor was affixed to an in-vacuum translator, allowing us to map the temperature gradients near the lattice-confined atoms (inset to Fig. 2c). During clock operations the mechanical translator was retracted to avoid interference with atoms (Fig. 2b). Systematic errors in both the readout of the sensors and their ability to determine the actual thermal distribution at the position of the atoms resulted in an overall uncertainty of 26.7 mK for the stated BBR temperature. Table 2 lists the sources of uncertainties for this temperature evaluation.

The atoms are influenced not only by the total integrated power of the BBR inside the chamber, known as the BBR static correction, but also the frequency-weighted spectrum of the radiation inside the chamber, known as the BBR dynamic correction. We constructed a ray-tracing model of our chamber to estimate the influence of temperature gradients throughout the vacuum chamber[24,25]. The model predicts the error incurred in the BBR dynamic correction by calculating the deviation from a perfect BBR spectrum for the temperature read out by the sensor[26] (see Methods). A perfect BBR environment corresponds to a spatially uniform temperature; deviations from such an environment cause a temperature gradient inside the chamber. As shown in Fig. 2c, components that couple strongly to the sensor, such as the large viewports, would need to deviate in temperature from the rest of the chamber by a significant amount (more than 10 K) for this error to reach the 1×10^{-18} level. Furthermore, within our BBR shielding box with small temperature gradients, the dominant emissivity-weighted solid angle of the vacuum viewports made the model's predictions for the dynamic BBR correction insensitive to the exact emissivity values. We ensured that the BBR shielding box is fully sealed from the outside environment, 'forbidding' the atoms to 'view' any highly emissive object with a temperature differing from the inside ambient temperature. This BBR shielding box also allowed the clock vacuum chamber to be insulated from room-temperature variations, because it reaches an equilibrium temperature of 301 K after two hours of clock operation.

Most systematics listed in Table 1 are rapidly measured through self-comparison with digital lock-in[6,27] (see Methods). Both SrI and SrII measure their systematics by modulating a particular physical parameter

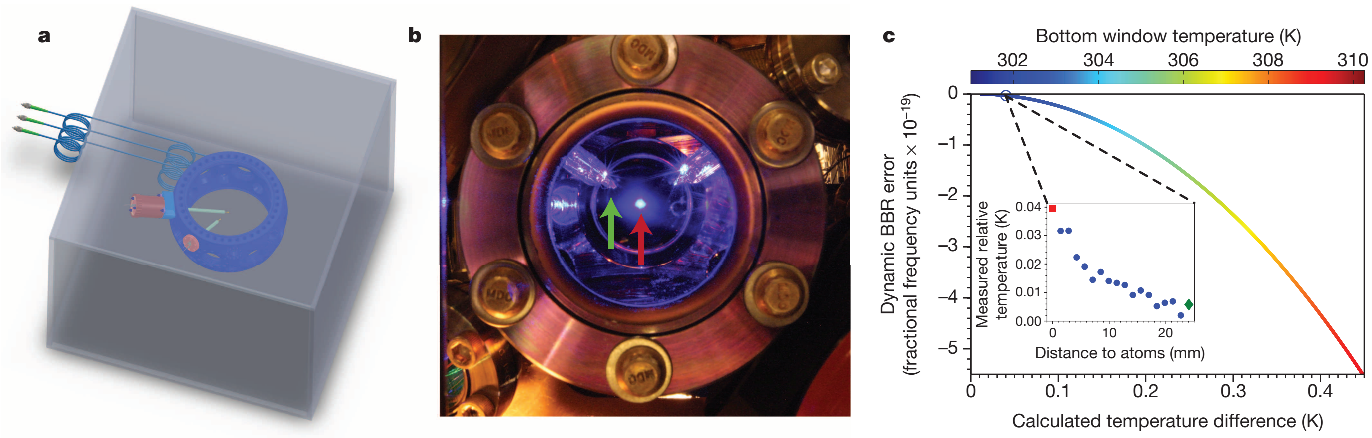


Figure 2 | Characterizing BBR effects on the $^1S_0$–$^3P_0$ transition. **a**, A three-dimensional model of the clock vacuum chamber. The sensor mounted on an in-vacuum translator is depicted in its fully extended mode of operation. The entire clock chamber resides inside a BBR shielding box with an equilibrium temperature of 301 K. **b**, A photograph of the two glass tubes surrounding the trapped ^{87}Sr atoms (red arrow). The movable sensor (green arrow) has been retracted for its normal operation. **c**, The error inherent in assuming a perfect BBR spectrum inside the vacuum chamber, based on a measurement of total BBR radiated power. Modelling all components of the chamber as 301 K and varying the bottom window temperature (shown in the top horizontal axis) shows that measuring the total radiative power is sufficient for our quoted BBR systematic uncertainty. The bottom horizontal axis displays the temperature difference between the atoms and the retracted sensor. The inset is a typical measured temperature difference inside the vacuum chamber referenced to the temperature of the retracted movable sensor at the beginning of the measurement. Green diamond, retracted position; red square, atomic position.

every two experimental cycles, with the clock laser serving as a stable reference with which to measure the related frequency shifts. For example, the atomic density shift was measured to high precision with this method. The SrII system is designed to reach a density shift uncertainty below 1×10^{-18} by using large lattice trapping volumes. To accommodate this, we used a Fabry–Perot buildup cavity to achieve a sufficiently deep lattice. This trap design increases the number of atoms loaded into the lattice at a decreased atomic density, allowing SrII to measure an already-reduced density shift to very high precision. Details of the SrI and SrII optical lattice trap geometries can be found in ref. 6 and in Methods.

Frequency shifts induced by the optical lattice potential must be understood and controlled at an extremely high level of precision, especially for optical lattices that trap weakly against gravity[28]. We used a variety of methods to stabilize the lattice scalar, vector and tensor Stark shifts. An 813-nm continuous-wave Ti:sapphire laser was used to create the lattice light for SrII. The clean spectrum of the solid-state laser has the advantage over a semiconductor tapered-amplifier-based lattice, where spontaneous noise pedestals might cause additional frequency shifts[9]. For SrI, we used a tapered-amplifier system, but we refined the output spectrum with a narrow-band interference filter and an optical filter cavity. To deal with potential residual shifts due to the tapered-amplifier noise pedestals, we regularly calibrated the lattice Stark shift for SrI. Both clocks stabilize their lattice laser frequencies to a Cs clock via a self-referenced Yb fibre comb, and their trapping light intensities were stabilized after being delivered to the atoms. The lattice vector shift was cancelled by alternately interrogating the $+9/2$ and $-9/2$ stretched nuclear spin states of the atom on successive experimental cycles, in addition to the use of linearly polarized lattice light[8,28]. This interrogation sequence also allowed cancellation of the first-order Zeeman shift. Rather than trying to separate the scalar and tensor shifts artificially, we treat them as a single effect in our measurement of the a.c. Stark shift[28]. (In reference to alternating (or direct) current, a.c. (or d.c.) is used to denote oscillatory (or static) fields and their effects.) We further minimized the tensor shift's sensitivity to the magnetic bias field B by setting the lattice polarization and the direction of B to be parallel. When modulating the intensity of the lattice, we did not identify any lattice shifts that are nonlinear in lattice intensity. Specifically, we eliminated systematic biases arising from differential atomic interaction shifts and optical spectrum shifts from the a.c. Stark effect. A Fisher test performed for various model shifts on an extensive set of data (shown in Fig. 3a) demonstrated that the lattice shift is consistent only with a linear model to within 1σ uncertainty (see Methods).

For SrII we also took extra care to minimize fluctuations in the magnetic-field-related lattice Stark effect and the second-order Zeeman shift[28]. To stabilize the magnetic field for our clock over long operational periods, we used the atoms themselves as a collocated magnetometer for the clock. Every two minutes during the clock operation, the computer-based frequency locking program was paused to interrogate unpolarized atomic samples under zero applied magnetic field. A drift in the background magnetic field resulted in a reduced excitation for the peak of an unpolarized line, because all ten nuclear spin states will experience different Zeeman shifts. Every time the magnetic field servo was activated, the program automatically dithered each pair of magnetic field compensation coils (along three orthogonal spatial directions) and optimized the current for each pair of coils. As shown in Fig. 3b, the magnetometer-based feedback loop not only keeps the field direction constant throughout the clock operation, but also automatically nulls the background field without an operator's intervention. For SrI, this procedure is unnecessary owing to its more stable magnetic field environment.

To push the systematic evaluation to the 10^{-18} level, we also needed to evaluate the d.c. electric-field-induced Stark effect, a frequency shift mechanism caused by patch charges immobilized in the vacuum chamber's fused silica viewports[29]. A pair of disk-shaped electrodes was placed near the two largest viewports in the system (separated along the vertical direction) and shifts were recorded as the electrode polarity was switched. Differences between the frequency shifts induced by oppositely charged electrodes indicate the presence of stray background electric fields, as shown in Fig. 3c. On first measuring the d.c. Stark

Table 2 | Uncertainties for the in-vacuum silicon diode thermometer

Corrections	ΔT (mK)	σ_T (mK)
Calibration (including self-heating)	0	16
Residual conduction	0	0.7
Temperature gradient	40	20
Lead resistance	7.7	1.5
Lattice light heating	−15	7.5
Totals	32.7	26.7

All uncertainties are quoted as 1σ standard errors. The absolute calibration of the silicon diode sensor (including the self-heating effect) was performed by the vendor (Lake Shore Cryotronics) and the calibration is traceable to the NIST blackbody radiation standard. We have evaluated corrections and their uncertainties for the operation of the sensors in our vacuum chamber, including the residual conduction by the mount, extra lead resistance and lattice light heating.

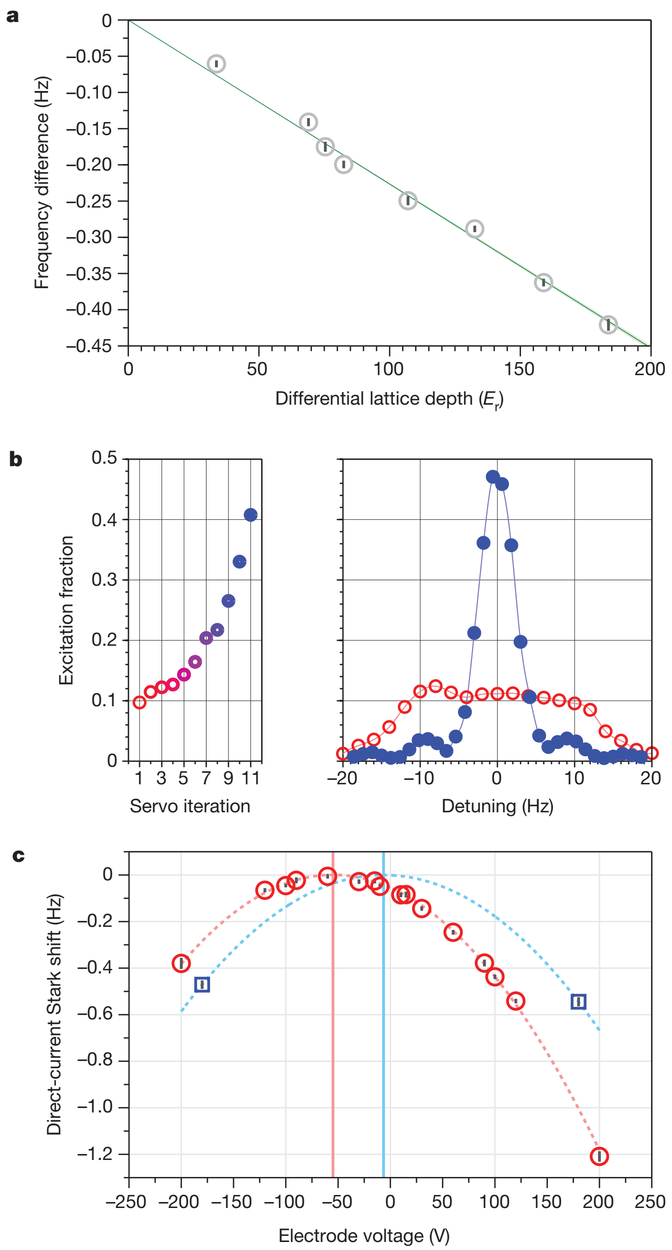


Figure 3 | Examples of systematic evaluations. **a**, To determine the lattice a.c. Stark effect accurately, a variety of lattice depths were used. This effect is depicted as a function of the differential lattice depth with binning chosen for figure clarity (average bin size 68 min, corresponding to an average of 1,600 points). Within our measurement precision, the best fit is a linear model. Grey circles denote mean frequency shifts and small solid black lines represent the 1σ standard errors inflated by the square root of the reduced chi-squared. The solid green line is a linear model and the light green patch represents the 1σ standard error for this model. **b**, Using the atomic cloud as a collocated magnetometer, a residual non-zero magnetic field is inferred via the peak excitation of an unpolarized Rabi lineshape. The left panel shows the servo action of zeroing the residual magnetic field. The right panel shows a clock transition lineshape for an unstabilized magnetic field (red open circles) and an improved lineshape under the stabilized magnetic field (blue filled circles). The red and blue solid lines are simply guides for the eye. **c**, Measurements of d.c. electric-field-induced Stark shift show a quadratic behaviour. The red circles show that a residual shift due to the stray d.c. field was -1.3×10^{-16}. The blue squares show a greatly reduced shift after purging the vacuum chamber with N_2 gas. Dashed lines show a quadratic fit to the data. Solid black lines represent the 1σ standard errors inflated by the square root of the reduced chi-squared. Solid red and blue vertical lines show the locations of zero net electric field.

effect on SrII, a residual, stable -1.3×10^{-16} shift was discovered. However, when the vacuum chamber was filled with clean nitrogen and then re-evacuated, we reduced the measurable d.c. Stark effect to $-1.6(1.0)\times10^{-18}$. To complete the full evaluation of the d.c. Stark effect, we performed similar measurements along the horizontal direction and determined its effect at $-1.9(1.9)\times10^{-18}$.

The outlook for optical lattice clocks is bright. We note that this is only the first systematic evaluation of a lattice clock enabled by the new generation of stable lasers, which led to clock stability near the quantum projection noise limit for 1,000 atoms. As laser stability continues to improve[30], Sr and other lattice clocks will increase their quantum-projection-noise-limited precision with larger numbers of atoms. Along with great advances in stability, the systematic uncertainty for such clocks will rapidly decrease owing to much reduced measurement times. Hence, the stability and total uncertainty of future lattice clocks will advance in lockstep. The techniques demonstrated here will allow for clock stability and total uncertainty below 1×10^{-18}. Such clocks will in turn push forward a broad range of quantum sensor technologies and facilitate a variety of fundamental physics tests.

METHODS SUMMARY

For both clocks, a few thousand ^{87}Sr atoms are laser cooled to around 3 μK and trapped in one-dimensional optical lattices near the magic wavelength (813 nm), with trap depths ranging from $40E_r$ to $300E_r$ (where E_r is the photon recoil energy). A thermal-noise-limited laser with a short-term stability of 1×10^{-16} (from 1 s to 1,000 s) interrogates the $^1S_0$–$^3P_0$ clock transition with Rabi spectroscopy for 160 ms. The clock comparison is normally operated in an asynchronous interrogation mode, where the two clock probe pulses are purposely non-overlapping in time[6]. Two independent acousto-optic modulators (AOMs) are used to correct the laser frequency to the SrI and SrII clock transitions. State detection of atomic ensembles is a destructive measurement that requires the repetition of the experimental cycle every 1.3 s. After each cycle, the frequency corrections, atom numbers and environmental temperatures for both systems are recorded and time stamped. For evaluation of the systematic uncertainty of the clock frequency, we focus on a few dominating effects such as the blackbody radiation, atomic interaction, lattice Stark shift and magnetic field, as well as a range of other sources of uncertainties such as the d.c. Stark shift, the clock laser a.c. Stark shift, line pulling and lattice tunnelling effects, AOM phase chirp, the second-order Doppler effect, background gas collisions and atomic servo errors. All experimentally measured quantities are treated with rigorous statistical analysis. Long time records of data are binned into various sizes of time windows, producing means and standard deviations of these bins along with the reduced chi-squared, $\sqrt{\chi^2_{\mathrm{reduced}}}$. When $\sqrt{\chi^2_{\mathrm{reduced}}}>1$, indicating overscatter in data, the smaller bins' standard deviations are scaled up to bring $\sqrt{\chi^2_{\mathrm{reduced}}}$ to 1, and the analysis to determine the systematic is repeated.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 September; accepted 4 December 2013.
Published online 22 January 2014.

1. Wineland, D. J. Nobel lecture: Superposition, entanglement, and raising Schrödinger's cat. *Rev. Mod. Phys.* **85,** 1103–1114 (2013).
2. Haroche, S. Nobel lecture: Controlling photons in a box and exploring the quantum to classical boundary. *Rev. Mod. Phys.* **85,** 1083–1102 (2013).
3. Chou, C. W., Hume, D. B., Koelemeij, J. C. J., Wineland, D. J. & Rosenband, T. Frequency comparison of two high-accuracy Al^+ optical clocks. *Phys. Rev. Lett.* **104,** 070802 (2010).
4. Huntemann, N. *et al.* High-accuracy optical clock based on the octupole transition in $^{171}Yb^+$. *Phys. Rev. Lett.* **108,** 090801 (2012).
5. Madej, A. A., Dubé, P., Zhou, Z., Bernard, J. E. & Gertsvolf, M. $^{88}Sr^+$ 445-THz single-ion reference at the 10^{-17} level via control and cancellation of systematic uncertainties and its measurement against the SI second. *Phys. Rev. Lett.* **109,** 203002 (2012).
6. Nicholson, T. L. *et al.* Comparison of two independent Sr optical clocks with 1×10^{-17} stability at 10^3s. *Phys. Rev. Lett.* **109,** 230801 (2012).
7. Hinkley, N. *et al.* An atomic clock with 10^{-18} instability. *Science* **341,** 1215–1218 (2013).
8. Ludlow, A. D. *et al.* Lattice clock at 1×10^{-16} fractional uncertainty by remote optical evaluation with a Ca clock. *Science* **319,** 1805–1808 (2008).

9. Le Targat, R. *et al.* Experimental realization of an optical second with strontium lattice clocks. *Nature Commun.* **4,** 2109, http://dx.doi.org/10.1038/ncomms3109 (2013).
10. Falke, S. *et al.* The ^{87}Sr optical frequency standard at PTB. *Metrologia* **48,** 399–407 (2011).
11. Bordé, C. J. Base units of the SI, fundamental constants and modern quantum physics. *Phil. Trans. R. Soc. A* **363,** 2177–2201 (2005).
12. Rosenband, T. *et al.* Frequency ratio of Al^{+} and Hg^{+} single-ion optical clocks; metrology at the 17th decimal place. *Science* **319,** 1808–1812 (2008).
13. Chou, C. W., Hume, D. B., Rosenband, T. & Wineland, D. J. Optical clocks and relativity. *Science* **329,** 1630–1633 (2010).
14. Martin, M. J. *et al.* A quantum many-body spin system in an optical lattice clock. *Science* **341,** 632–636 (2013).
15. Ye, J., Kimble, H. J. & Katori, H. Quantum state engineering and precision metrology using state-insensitive light traps. *Science* **320,** 1734–1738 (2008).
16. Takamoto, M., Hong, F.-L., Higashi, R. & Katori, H. An optical lattice clock. *Nature* **435,** 321–324 (2005).
17. Kessler, T. *et al.* A sub-40-mHz-linewidth laser based on a silicon single-crystal optical cavity. *Nature Photon.* **6,** 687–692 (2012).
18. Bishof, M., Zhang, X., Martin, M. J. & Ye, J. Optical spectrum analyzer with quantum limited noise floor. *Phys. Rev. Lett.* **111,** 093604 (2013).
19. Campbell, G. K. *et al.* Probing interactions between ultracold fermions. *Science* **324,** 360–363 (2009).
20. Swallows, M. D. *et al.* Suppression of collisional shifts in a strongly interacting lattice clock. *Science* **331,** 1043–1046 (2011).
21. Lemke, N. D. *et al.* p-Wave cold collisions in an optical lattice clock. *Phys. Rev. Lett.* **107,** 103902 (2011).
22. Middelmann, T., Falke, S., Lisdat, C. & Sterr, U. High accuracy correction of blackbody radiation shift in an optical lattice clock. *Phys. Rev. Lett.* **109,** 263004 (2012).
23. Boyd, M. M. *et al.* Optical atomic coherence at the 1-second time scale. *Science* **314,** 1430–1433 (2006).
24. Chandos, R. J. & Chandos, R. E. Radiometric properties of isothermal, diffuse wall cavity sources. *Appl. Opt.* **13,** 2142–2152 (1974).
25. Yasuda, M. & Katori, H. Lifetime measurement of the $^{3}P_{2}$ metastable state of strontium atoms. *Phys. Rev. Lett.* **92,** 153004 (2004).
26. Middelmann, T. *et al.* Tackling the blackbody shift in a strontium optical lattice clock. *IEEE Trans. Instrum. Meas.* **60,** 2550–2557 (2011).
27. Boyd, M. *et al.* ^{87}Sr lattice clock with inaccuracy below 10^{-15}. *Phys. Rev. Lett.* **98,** 083002 (2007).
28. Westergaard, P. G. *et al.* Lattice-induced frequency shifts in Sr optical lattice clocks at the 10^{-17} level. *Phys. Rev. Lett.* **106,** 210801 (2011).
29. Lodewyck, J., Zawada, M., Lorini, L., Gurov, M. & Lemonde, P. Observation and cancellation of a perturbing dc stark shift in strontium optical lattice clocks. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **59,** 411–415 (2012).
30. Cole, G. D., Zhang, W., Martin, M. J., Ye, J. & Aspelmeyer, M. Tenfold reduction of Brownian noise in high-reflectivity optical coatings. *Nature Photon.* **7,** 644–650 (2013).

Acknowledgements We thank M. Martin, M. Swallows, E. Arimondo, J. L. Hall, T. Pfau, and W. D. Phillips for discussions and H. Green for technical assistance. This research is supported by the National Institute of Standards and Technology, the Defense Advanced Research Projects Agency's QuASAR Program, and the NSF PFC. M.B. acknowledges support from the National Defense Science and Engineering Graduate fellowship programme. S.L.C. and M.B. acknowledge support from the NSF Graduate Fellowship. Any mention of commercial products does not constitute an endorsement by NIST.

Author Contributions B.J.B., T.L.N., J.R.W., S.L.C., M.B., X.Z., W.Z., S.L.B. and J.Y. conceived, designed and carried out the experiments mentioned in this manuscript. All authors discussed the results and contributed to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.Y. (ye@jila.colorado.edu).

METHODS

Atomic sample preparation. For both clocks, up to a few thousand ^{87}Sr atoms are laser cooled to a few microkelvin and trapped in one-dimensional optical lattices near the magic wavelength (813 nm), with trap depths ranging from $40E_r$ to $300E_r$. Here E_r is the photon recoil energy. A thermal-noise-limited laser with a short-term stability of 1×10^{-16} (from 1 s to 1,000 s) interrogates the $^1S_0$–$^3P_0$ clock transition with Rabi spectroscopy over a 160-ms probe time. We allow a sufficient time for transient perturbations in the system to decay before we interrogate the clock transition. We normally operate the clock comparison in an asynchronous interrogation mode, where the two clock probe pulses are purposely non-overlapping in time[6]. Two independent frequency shifters (acousto-optic modulators (AOMs)) are used to correct the laser frequency to the SrI and SrII clock transitions. State detection of both atomic ensembles is a destructive measurement that requires the repetition of the experimental cycle every 1.3 s. After each cycle, the frequency corrections, atom numbers, and environmental temperatures for both systems are recorded and time stamped for comparison and post-processing.

Statistical methods for data analysis. For all systematic measurements, residual non-white noise introduces overscatter in the data. Following our previously reported procedure[20], the data are first binned into smaller chunks, the means and standard deviations of these bins are determined, and a reduced chi-squared, $\sqrt{\chi^2_{\mathrm{reduced}}}$, is obtained. In instances where $\sqrt{\chi^2_{\mathrm{reduced}}} > 1$, indicating overscatter in the data, the smaller bins' standard deviations are inflated to bring $\sqrt{\chi^2_{\mathrm{reduced}}}$ to 1, and the analysis to determine the systematic is repeated. This conservative approach is applied to all measurements in this Letter unless otherwise noted.

Previously, the most comprehensive systematic evaluations of optical lattice clocks with total systematic uncertainty better than that of Cs were reported in refs 8–10 and 31. During the production of this manuscript, another systematic evaluation of Sr from the Physikalisch-Technische Bundesanstalt group has been released[32]. Below, we provide a detailed discussion of the systematics we evaluated in this work.

Blackbody radiation shifts. The blackbody radiation shift is determined by $\Delta\nu_{\mathrm{BBR}} = -2.13023(T/300)^4 - 0.1484(T/300)^6$, where T is the temperature in K and $\Delta\nu_{\mathrm{BBR}}$ has units of Hz. The T^4 term is known as the static shift, and the T^6 term is called the dynamic shift. To ascertain the radiative temperature experienced by the atoms, silicon diode temperature sensors were installed in SrII. Silicon diode sensors are used for their ease of calibration (because their forward voltage drop is linear in temperature) and their suitability for vacuum baking[33]. We investigate the thermalization of the probes by modelling their heat transfer. For the small thermal gradients measured around our chamber, the sensors give a good measurement of the integrated BBR spectrum, which is proportional to the static BBR shift experienced by the atoms.

The dynamic shift, which depends on the frequency-weighted spectrum of radiation experienced by the atoms, is calculated using the temperature read out by the sensor. The T^6 coefficient associated with this shift was chosen to be a simple mean between the two most recent publications and the uncertainty in this coefficient was chosen to be their difference[22,34]. This coefficient uncertainty is the dominant BBR uncertainty. To understand the error we accrue by calculating the dynamic shift with the sensor reading, a ray-tracing model of our chamber was constructed. Refraining from using a Monte Carlo population of rays, we used Hammersley boundary points to construct the ray population in a controlled, repeatable and processor-efficient manner. Effective emissivity-weighted solid angles are then tabulated according to whatever position inside the main chamber the end user requires. By keeping track of the approximate blackbody radiation spectrum at any point in the chamber and using the relevant Einstein A coefficients of Sr in ref. 22, the error associated with assuming a perfect BBR spectrum was determined. The model's results, as seen in Fig. 2c, show that this error is well below our quoted systematic uncertainty for BBR under small temperature variations around our chamber.

Within our BBR shielding box with small temperature gradients, the dominant emissivity-weighted solid angle of the vacuum viewports makes the model's predictions for the dynamic BBR correction insensitive to the exact emissivity values. For example, for small temperature differentials (3 K) across our chamber, the emissivity of the metal would need to change by a factor of 20 to introduce a 1×10^{-18} error into the model's predictions.

However, much hotter components that couple more strongly to the atoms, such as the heated Zeeman slower window, introduce a larger error in the deviation from a perfect BBR spectrum. For the systematic uncertainty evaluation, the operation of the Sr optical lattice clock can be performed without heating the Zeeman slower window for a limited amount of time. Even with the Zeeman slower window heated, its contribution to total uncertainty is below 1.2×10^{-18}. For future experiments that require total uncertainty below 1×10^{-18}, we can simply add mechanical shutters that obscure the atoms' view of all hot elements in the system. SrI's temperature is a simple weighted mean, based on a set of temperature measurements made at various points on the chamber. The weights for this mean are derived from the emissivity-weighted solid angles the atoms experience from various components. We conservatively quote the full range of temperature across the SrI chamber (0.7 K) as the SrI temperature uncertainty.

Atomic density shifts. For both SrI and SrII, the atomic density shift is measured via self-comparison by modulating the lattice trapped atom number during subsequent measurements of the line centre. The fast modulation timescale makes these measurements immune to long-term atom number drifts. Extrapolation of the density shift from changes in atomic density due to changes in the lattice trapping potential is used only when trap frequencies have changed by less than 10%. Any additional error from this extrapolation is included in the final quoted uncertainty.

Magnetic field effects. The first-order Zeeman shifts are cancelled by alternately interrogating opposite nuclear spin stretched states. After each line centre acquisition of the clock laser, the quoted Sr frequency is the running average between the two previous measurements. Key to the efficacy of this scheme is that changes in the magnetic field over two line centre acquisitions are either negligible or averaged away. The residual first-order Zeeman shift is calculated by examining the overall drift of the magnetic field splitting between the two stretched nuclear spin states, and extrapolating the inherent shift this drift would induce.

The second-order Zeeman shifts, caused by the presence of an applied bias and a residual magnetic field during the clock interrogation sequence, are subtracted off point by point. The determination of the second-order Zeeman shift coefficient is made via a fast modulation experiment whereby four digital locks are modulated in sequence: two locks at a high magnetic field, one for a measurement of each stretched spin component, and two locks at a low magnetic field. This experiment was performed for stretched states' splitting ranging from 300 Hz to 1,200 Hz. To fit this data, a quadratic function, cS^2, was used. Here c was found to be $-0.248(2) \times 10^{-6}\,\mathrm{Hz}^{-1}$, and S is the measured stretched state splitting in Hz. Extra care was taken to ensure that modulation of the applied bias field did not cause any rotation of the polarization axis, which would induce an unwanted differential lattice tensor shift. The second-order Zeeman shift uncertainty is quoted for a 500-Hz splitting between the $m_f = \pm 9/2$ stretched states, which allows for a relatively strong bias magnetic field (about 50 μT) to be applied to the atoms during clock operation.

During the course of the previous measurement, instabilities in the magnetic field environment of SrII were discovered. Active control of the magnetic field was implemented to combat this. As detailed in the main text, the peak excitation of an unpolarized line is used as a measure of the residual magnetic field in the system. Once the clock has entered into the servo routine, it takes three peak excitation measurements for each coil at different currents ($I_0 - \Delta I$, I_0, $I_0 + \Delta I$, where I_0 is the current of the previous iteration and ΔI is a small trial step). In situations where the residual field is far from zero, the software steps the current by a fixed amount in the direction indicated by the increasing excitation. When the field is near zero, a parabola was fitted between the three measurements and the current was stepped to the fitted value filtered via a low-pass finite impulse response filter. On stopping the servo routine and measuring the unpolarized line, the residual magnetic field wander was measured to be less than 0.3 μT.

Lattice Stark effects. To achieve the low uncertainties reported for SrII, a variety of new techniques were used to minimize any systematic errors in the measurement of the lattice stark effects. The intensity servos were implemented with a liquid-crystal waveplate for SrI and an acousto-optic modulator for SrII. Trap frequency measurements were performed at each measured a.c. Stark point via a high-resolution sideband scan. As explained in ref. 35, radial motion only brings the sidebands closer to the carrier, so the true longitudinal sideband frequency is found by looking at the farthest edge of the blue sideband. This edge corresponds to the contributions of atoms that are distributed at the centre of the Gaussian profile for the lattice beam. A tangent line was fitted to the inflection point of the Lorentzian, and then the Lorentzian centre frequency could be extrapolated from the tangent x-intercept and the Lorentzian linewidth determined by the Fourier-limited linewidth of the clock laser scan. A scan of the carrier was taken simultaneously with these high-resolution sideband scans, and a fit to the carrier was used to determine the Lorentzian linewidth. Once the longitudinal trap frequency was determined, the trap depth could be extracted using the procedures outlined in ref. 35. Alternating between four atomic servos, with the magnetic field control activated, we measured differential shifts between a variety of high and low lattice depths from $87E_r$ to $300E_r$ for the SrII apparatus. To minimize systematic uncertainties caused by differential atomic density shifts, both high and low lattice depths were operated with an absolute density shift below 1×10^{-17}. Atomic interaction shifts follow a power-law behaviour in the trap frequency, and must be taken into account especially at very high lattice depths. Atom numbers were chosen to provide similar density shifts for both high and low lattice depths. Furthermore, individual measurements for particular values of lattice depth difference were performed at 1×10^{-17} statistical uncertainty. Many measurements at different lattice depths were needed to achieve the low uncertainty reported here (see Fig. 3a). Raw measurement data are binned in sets of 30 to 75 points per bin

according to the procedure outlined in the 'Statistical methods for data analysis' section. To decrease the sensitivity to bin choice, the final uncertainties for the fit parameter are simple means of the fit parameter errors determined for all bin sizes. Model fits for both a hyperpolarizability and an E2/M1 contribution to the measured stark shift revealed no statistically significant contribution, based on a Fisher test with a 1σ threshold. Even at this measurement precision, the small magnitude of these shifts allowed their contributions to be included in the linear fit. Our fitted hyperpolarizability, 0.48(47) μHz/E_r^2, is not inconsistent with previously reported coefficients[9] of 0.45(10) μHz/E_r^2. However, our data, as shown by the Fisher test, support a linear fit only. We thus list only a single overall systematic uncertainty for the lattice a.c. Stark shift, treating the entire data set in the most statistically consistent manner. We note that using a prior reported hyperpolarizability coefficient[9], our overall a.c. Stark uncertainty would change only slightly from 3.7×10^{-18} to 4.1×10^{-18}. However, in that previous measurement, to gain a very large lever arm for the measurement of the hyperpolarizability, very high (5,000E_r) lattice depths were used[9], but the atomic density effects in such tight traps were not considered. In summary, although our data are not inconsistent with the hyperpolarizability measurements previously reported[9], this work is not an independent verification of the previous measurement. Our data alone do not support a statistically significant nonlinearity, and the extracted hyperpolarizability would have a higher uncertainty than the previous measurement.

Although, to first order, lattice vector shifts are cancelled by alternately measuring opposite nuclear spin stretched states, a residual lattice vector shift can cause systematic shifts due to its convolution with the second-order Zeeman shift. A lattice vector shift will cause an overall widening of the shift between opposite nuclear spin states, mimicking a magnetic field. As a conservative estimate, we included the effect of a 100-mHz residual lattice vector splitting. To calculate how this will affect the uncertainty related to the second-order Zeeman effect, we included this 100 mHz as an error to S, as defined in the 'Magnetic field effects' section.

Miscellaneous shifts. Shifts from background gas collisions were estimated using the methods described in ref. 36. Differential C_6 coefficients for the Sr $^1S_0$ and $^3P_0$ states[37] for their resonant dipole interactions were scaled to the Cs ground-state–ground-state C_6 coefficients. By far the largest residual gas in our ultrahigh-vacuum, oven-loaded system is hydrogen. Both the Sr $^1S_0$-H_2 C_6 coefficient and the Sr $^3P_0$-H_2 C_6 coefficient were then estimated by scaling with respect to the non-resonant dipole Cs–H_2 C_6 coefficient. Atomic trapping lifetimes of about 1 s (SrII) and about 8 s (SrI), average excitations during the detuned Rabi pulse and interrogation times of 160 ms were combined to estimate the background gas collisional shift uncertainty. No shift correction is quoted, and we provide only an upper bound of this uncertainty.

Using the measured temperature of around 3 μK, we calculate an average total velocity of 3 cm s^{-1}. A Taylor expansion of the full relativistic Doppler shift has second-order terms from both longitudinal and transverse motion. Overall, the second-order Doppler effect results in a fractional shift less than 10^{-20}.

Line pulling can be caused by a variety of effects. These effects include a slight ellipticity in the clock laser polarization, imperfect optical pumping (<5% population in neighbouring nuclear spin states), and clock-laser-induced tunnelling (no signal was visible for tunnelling-induced sidebands). They can be modelled as a deformation of a perfect Rabi lineshape. The spectral narrowness of the 5-Hz Fourier-limited linewidth with which all data was taken in this work greatly reduced the effect of any possible line pulling.

The only exception to taking data with spectrally narrow features was the investigation of the a.c. Stark shift induced by the clock laser itself. This systematic was evaluated by measuring the frequency difference between our clock transition interrogated with 50-ms and 200-ms π pulses using our fast modulation technique. For this measurement, the clock was run with our largest possible bias magnetic field to avoid any residual line pulling effects during the 50-ms clock interrogation. Including the errors in determining our π pulse exactly, we estimate a 1.3×10^{-18} uncertainty in the probe beam a.c. Stark shift for SrII's normal clock operation.

An AOM is used to scan the ^{87}Sr clock transition and to shape our laser pulse. When the clock pulse is switched on, phase transients originating from this AOM can cause a measured frequency shift. We compared light from a diffracted order of this AOM with the zeroth-order light using a digital phase detector. After we calibrated and removed the effect of the detector's phase transients from our data, the observed effects, when convolved with the sensitivity function of Rabi spectroscopy, resulted in a shift almost consistent with zero[38].

Servo error was determined by combining many hours of lock data and measuring whether there is a systematic bias to the in-loop error signal. Any bias measured was transformed to a frequency shift and uncertainty by modelling a perfect Rabi lineshape with the contrast and pulse area under which the data was taken. This allowed us to transform error signal bias to frequency shifts.

Frequency comparison. After the data had been post-processed by each individual strontium team, time-stamped and corrected frequencies were shared. Although the overall systematic uncertainty of the comparison is 5.4×10^{-17}, as a consistency check for the comparison a variety of methods were used to show the agreement of the two clocks within this confidence interval. A simple mean of all the data gives the difference between the two clocks to be -2.4×10^{-17}. Binning the data in small chunks, of approximately one minute per data point (as in the top panel of Fig. 1b) gives agreement of $-2.8(2) \times 10^{-17}$. The uncertainty on this number has been inflated by $\sqrt{\chi^2_{\text{reduced}}}$, because $\sqrt{\chi^2_{\text{reduced}}} = 3.5$ denotes overscatter in the data. Binning the data in 30-min chunks (as in the bottom panel of Fig. 1b) clearly shows that there are systematic fluctuations still present in the comparison, with $\sqrt{\chi^2_{\text{reduced}}} = 10.5$ and an agreement of $-2.7(5) \times 10^{-17}$. Again, this uncertainty is inflated by $\sqrt{\chi^2_{\text{reduced}}}$. The greater overscatter in the data at longer timescales is probably caused by imprecise knowledge of the BBR environment for SrI, which allows for fluctuations within the 1σ comparison uncertainty.

The final systematic uncertainty used in the comparison is quoted under the running conditions of the two strontium systems during the comparison, and not their final best achieved total uncertainties. Furthermore, the height difference (10 cm) between the two atomic clouds, resulting in a 1.0×10^{-17} gravitational redshift, was included in the comparison but was not relevant for Table 1.

31. Campbell, G. K. *et al.* The absolute frequency of the ^{87}Sr optical clock transition. *Metrologia* **45,** 539–548 (2008).
32. Falke, S. *et al.* A strontium lattice clock with 3×10^{-17} inaccuracy and its frequency. Preprint at http://arxiv.org/abs/1312.3419 (2013).
33. McNamara, A. G. Semiconductor diodes and transistors as electrical thermometers. *Rev. Sci. Instrum.* **33,** 330–333 (1962).
34. Safronova, M. S., Porsev, S. G., Safronova, U. I., Kozlov, M. G. & Clark, C. W. Blackbody-radiation shift in the Sr optical atomic clock. *Phys. Rev. A* **87,** 012509 (2013).
35. Blatt, S. *et al.* Rabi spectroscopy and excitation inhomogeneity in a 1D optical lattice. *Phys. Rev. A* **80,** 052703 (2009).
36. Gibble, K. Scattering of cold-atom coherences by hot atoms: frequency shifts from background-gas collisions. *Phys. Rev. Lett.* **110,** 180802 (2013).
37. Santra, R., Christ, K. & Greene, C. Properties of metastable alkaline-earth-metal atoms calculated using an accurate effective core potential. *Phys. Rev. A* **69,** 042510 (2004).
38. Falke, S., Misera, M., Sterr, U. & Lisdat, C. Delivering pulsed and phase stable light to atoms of an optical clock. *Appl. Phys. B* **107,** 301–311 (2012).

LETTER

doi:10.1038/nature12957

Drought sensitivity of Amazonian carbon balance revealed by atmospheric measurements

L. V. Gatti[1]*, M. Gloor[2]*, J. B. Miller[3,4]*, C. E. Doughty[5], Y. Malhi[5], L. G. Domingues[1], L. S. Basso[1], A. Martinewski[1], C. S. C. Correia[1], V. F. Borges[1], S. Freitas[6], R. Braz[6], L. O. Anderson[5,7], H. Rocha[8], J. Grace[9], O. L. Phillips[2] & J. Lloyd[10,11]

Feedbacks between land carbon pools and climate provide one of the largest sources of uncertainty in our predictions of global climate[1,2]. Estimates of the sensitivity of the terrestrial carbon budget to climate anomalies in the tropics and the identification of the mechanisms responsible for feedback effects remain uncertain[3,4]. The Amazon basin stores a vast amount of carbon[5], and has experienced increasingly higher temperatures and more frequent floods and droughts over the past two decades[6]. Here we report seasonal and annual carbon balances across the Amazon basin, based on carbon dioxide and carbon monoxide measurements for the anomalously dry and wet years 2010 and 2011, respectively. We find that the Amazon basin lost 0.48 ± 0.18 petagrams of carbon per year ($Pg\,C\,yr^{-1}$) during the dry year but was carbon neutral (0.06 ± 0.1 $Pg\,C\,yr^{-1}$) during the wet year. Taking into account carbon losses from fire by using carbon monoxide measurements, we derived the basin net biome exchange (that is, the carbon flux between the non-burned forest and the atmosphere) revealing that during the dry year, vegetation was carbon neutral. During the wet year, vegetation was a net carbon sink of 0.25 ± 0.14 $Pg\,C\,yr^{-1}$, which is roughly consistent with the mean long-term intact-forest biomass sink of 0.39 ± 0.10 $Pg\,C\,yr^{-1}$ previously estimated from forest censuses[7]. Observations from Amazonian forest plots suggest the suppression of photosynthesis during drought as the primary cause for the 2010 sink neutralization. Overall, our results suggest that moisture has an important role in determining the Amazonian carbon balance. If the recent trend of increasing precipitation extremes persists[6], the Amazon may become an increasing carbon source as a result of both emissions from fires and the suppression of net biome exchange by drought.

To observe the state, changes and climate sensitivity of the Amazon carbon pools we initiated a lower-troposphere greenhouse-gas sampling programme over the Amazon basin in 2010, measuring bi-weekly vertical profiles of carbon dioxide (CO_2), sulphur hexafluoride (SF_6) and carbon monoxide (CO) from just above the forest canopy to 4.4 km above sea level (a.s.l.) at four locations spread across the basin (Fig. 1). Repeated measurements of the CO_2 mole fraction in the low to mid-troposphere have the ability to constrain surface CO_2 fluxes at regional scales (about 10^5–10^6 km^2) including all known and unknown processes. This is in contrast to small temporal[8,9] and spatial[10,11] scale atmospheric approaches, which need substantial and difficult-to-verify assumptions to scale up; it is also in contrast to basin-scale surface-based studies, which include only a subset of relevant processes[3,12,13].

Our selection of sites reflects the dominant mode of horizontal air flow at mid- to low-troposphere altitudes across the Amazon basin, with air entering the basin from the equatorial Atlantic Ocean, sweeping over the tropical forested region towards the Andes and turning southwards and back to the Atlantic (Fig. 1). Air at the end-of-the-basin sites Tabatinga (TAB) and Rio Branco (RBA) is thus exposed to carbon fluxes from a large fraction of the basin's rainforest vegetation. Flux signatures

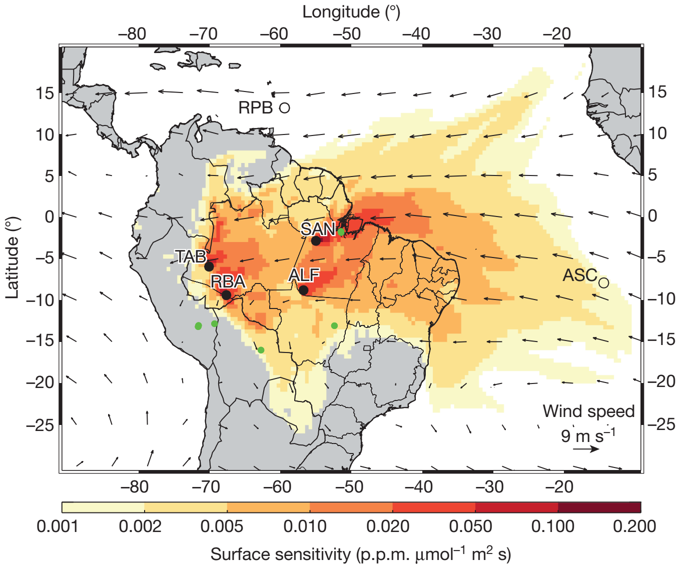


Figure 1 | Station's region of influence ('footprint'). The combined sensitivity of all observed atmospheric CO_2 concentrations to surface fluxes (that is, measurement 'footprints') is shown for the four sites TAB, RBA, SAN and ALF (solid black dots). Sensitivity is given in units of concentration (p.p.m.) per unit flux ($\mu mol\,m^{-2}\,s^{-1}$). As seen in Extended Data Fig. 6a, footprints from the four sites overlap substantially. Footprints are calculated at 0.5-degree resolution using ensembles of stochastically generated back trajectories using the FLEXPART Lagrangian particle dispersion model and then calculating the residence times of these back trajectories in the 100 m layer above the surface. Values above 0.001 p.p.m. $\mu mol^{-1}\,m^{-2}\,s^{-1}$ comprise 97% of the land surface signal and values above 0.01 p.p.m. $\mu mol^{-1}\,m^{-2}\,s^{-1}$ comprise 50% of the land surface signal; thus apparently small values are still important because they occupy a large area. Black arrows represent average climatological wind speed and direction in June, July and August (from the National Centers for Environmental Prediction (NCEP); http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html) averaged between the surface and 600 mbar. Open symbols (RPB and ASC) represent the NOAA tropical Atlantic sites used to define the background concentrations of CO_2, CO and SF_6 coming into the Amazon basin. Solid green dots indicate the locations of forest plot clusters where long-term biomass gains and respiration have been observed.

[1]Instituto de Pesquisas Energéticas e Nucleares (IPEN)–Comissao Nacional de Energia Nuclear (CNEN)–Atmospheric Chemistry Laboratory, 2242 Avenida Professor Lineu Prestes, Cidade Universitaria, Sao Paulo CEP 05508-000, Brazil. [2]School of Geography, University of Leeds, Woodhouse Lane, Leeds LS9 2JT, UK. [3]Global Monitoring Division, Earth System Research Laboratory, National Oceanic and Atmospheric Administration, 325 Broadway, Boulder, Colorado 80305, USA. [4]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, Colorado 80309, USA. [5]Environmental Change Institute, School of Geography and the Environment, University of Oxford, South Parks Road, Oxford OX1 3QY, UK. [6]Center for Weather Forecasts and Climate Studies, Instituto Nacional de Pesquisas Espaciais (INPE), Rodovia Dutra, km 39, Cachoeira Paulista CEP 12630-000, Brazil. [7]Remote Sensing Division, INPE (National Institute for Space Research), 1758 Avenida dos Astronautas, São José dos Campos CEP 12227-010, Brazil. [8]Departamento de Ciencias Atmosfericas/Instituto de Astronomia e Geofisica (IAG)/Universidade de Sao Paulo, 1226 Rua do Matao, Cidade Universitaria, Sao Paulo CEP 05508-090, Brazil. [9]Crew Building, The King's Buildings, West Mains Road, Edinburgh EH9 3JN, UK. [10]School of Tropical and Marine Biology and Centre for Terrestrial Environmental and Sustainability Sciences, James Cook University, Cairns 4870, Queensland, Australia. [11]Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot SL5 7PY, Berkshire, UK.
*These authors contributed equally to this work.

in air at the other two sites, Alta Floresta (ALF) and Santarém (SAN) are not only from forests but also from savanna and agricultural land. Our measurements represent the first network of ongoing, well-calibrated CO_2 measurements over a large stretch of tropical land. Such measurements are vital, because the near-absence of CO_2 measurements sensitive to the tropical biosphere is the underlying cause of the large uncertainties in net flux estimates for tropical regions obtained by inverse modelling of atmospheric CO_2 (refs 14 and 15).

Fortuitously, the two years of atmospheric observations reported here are for an unusually dry year followed by a wet one (Fig. 2 and Extended Data Fig. 1a, b). Our measurements thus document the sensitivity of Amazon basin carbon pools to the effect of drought. The reasons for the dry conditions in 2010 were twofold. For the first three months an El Niño episode caused dry conditions in the north and centre of the Amazon basin, whereas during the second half of the year a positive North Atlantic sea surface temperature anomaly locked the inter-tropical convergence zone (where the northeast and southeast trade winds converge) into a position that was more northerly than usual. This caused enhanced and prolonged dry conditions in the southern areas of the Amazon basin (Extended Data Fig. 1a, b). A simple diagnostic of the stress on vegetation exerted by the negative precipitation anomalies is the climatological water deficit (CWD[16]; see Methods and Fig. 2), in which in 2010 large negative anomalies occurred for the northwestern basin. This is consistent with river discharge records[17]. Lesser negative anomalies in the northeastern basin were caused by early-year negative precipitation anomalies and the central–eastern and southern parts of the Amazon basin ('the arc of deforestation') had anomalies caused by low precipitation during the third quarter of the year. Monthly mean temperatures (Extended Data Fig. 1c, d) in 2010 were higher than average in every month, with especially large anomalies in February/March and August/September. These mirror the periods of greatest negative precipitation anomalies. Warmer than average temperatures (with respect to the last three decades) were also observed for every month of 2011, but 2011 was also an unusually wet year (Extended Data Fig. 1a, b). As shown below, observed basin-wide carbon flux variations for 2010 and 2011 reflect these temporal precipitation patterns.

To isolate the contribution of Amazon terrestrial carbon sources and sinks to the atmospheric CO_2 profiles, we first subtract a scalar background mole fraction from each of the observed profiles. This background represents the composition of air entering the Amazon basin from the Atlantic and is estimated as a weighted average of CO_2 at Ascension Island (ASC) and Ragged Point, Barbados (RPB) using a linear mixing model based on ASC and RPB SF_6 with weights determined from SF_6 measured at the site[18–20] (Methods). SF_6 is well suited for this purpose (that is, to estimate the fractional contributions of Northern and Southern Hemispheric air entering the basin) because it has a large inter-hemispheric difference (Extended Data Figure 8) and virtually no Amazonian emissions[21].

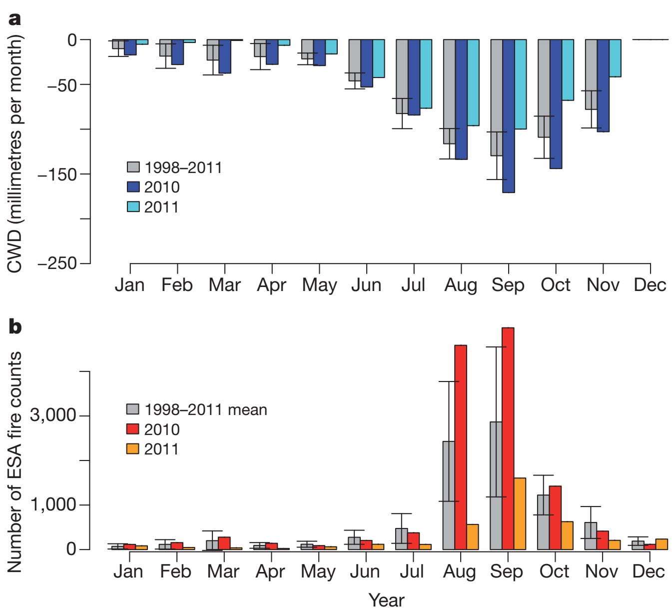


Figure 2 | Climatological water deficit. **a**, Basin-wide averages and standard deviation of CWD, based on the Tropical Rainfall Measuring Mission[28]. **b**, Fire counts based on European Space Agency (ESA; http://due.esrin.esa.int/wfa/) fire count data[29] for 2010, 2011 and 1998–2011, respectively.

Carbon sources and sinks reveal themselves in the referenced profiles $\Delta X = X_{site} - X_{bg}$ as mole fraction enhancements and depletions, where X is the mole fraction of CO_2 or CO, for site and background. The enhancements and depletions are generally confined to the lowermost 2 km or so of the profiles (Fig. 3). For ΔCO_2 (Fig. 3a–d), there is a strong tendency towards surface enhancements during the dry season, although both lower-troposphere depletions and enhancements can be observed at any time of the year. Vertical profiles of ΔCO show very large enhancements above the Atlantic background in the dry season, persisting into the free troposphere (Fig. 3e–h and Extended Data Fig. 2). CO is a product of incomplete combustion and in the Amazon it reflects a contribution to CO_2 enhancements from biomass burning. This is confirmed by calculated air-mass back-trajectories intersecting satellite-sensed fire hotspots (Extended Data Fig. 3) and by our observed CO:CO_2 ratios, which are typical for those from tropical forest fires (Methods).

From the profiles of ΔX we estimate fluxes by dividing them by the air-mass travel time t from the coast to the sampling site and integrating from the surface (0 km above ground level, a.g.l.) to 4.4 km a.s.l. determined by air-mass back-trajectories calculated separately for each of (typically) 12 air samples per profile[18–20] to obtain:

$$F_X = \int_{z\,=\,0\text{ km (a.g.l.)}}^{4.4\text{ km a.s.l.}} \frac{\Delta X}{t(z)} \mathrm{d}z \qquad (1)$$

Using measured CO:CO_2 emission ratios, $r^{bb}_{CO_2:CO}$ (refs 9 and 20), we further estimate the biomass burning contribution ($F^{bb}_{CO_2}$) to the net carbon flux using:

$$F^{bb}_{CO_2} = r^{bb}_{CO:CO_2}\left(F_{CO} - F^{bio}_{CO}\right) \qquad (2)$$

where F^{bio}_{CO} is the stable (background) value of F_{CO} during the wet season[20], reflecting direct plant and soil CO emissions as well as production from rapid oxidation of biogenic volatile organic compounds[22]. The non-fire net biome exchange (NBE) flux $F^{NBE}_{CO_2}$ is then given by:

$$F^{NBE}_{CO_2} = F^{total}_{CO_2} - F^{bb}_{CO_2} \qquad (3)$$

Our flux calculations (Fig. 4 and Table 1) reveal basin-wide average total fluxes of 0.19 ± 0.07 g C m^{-2} d^{-1} in 2010 and 0.02 ± 0.04 g C m^{-2} d^{-1} in 2011. Riverine carbon outgassing[13] is included in these fluxes but contributes minimally because the riverine organic carbon loop is very nearly closed within the Amazon basin[23], and fossil fuel emissions in the basin are negligibly small (<0.02 Pg C yr^{-1}; see Methods). Flux uncertainties presented in Fig. 4 and Table 1 may be underestimates because of losses of surface signal above 4.4 km caused by convective processes not captured by our extrapolation technique (Extended Data Table 1a). Our imperfect knowledge of convection and the difficulty of measuring CO_2 in the upper troposphere hamper quantification of these errors.

Using a basin area of 6.77×10^6 km^2 we calculate a source to the atmosphere of 0.48 ± 0.18 Pg C in 2010. In contrast, 2011 displayed an approximately neutral carbon balance (0.06 ± 0.10 Pg C yr^{-1}). In 2010, we calculate carbon losses due to fires of 0.51 ± 0.12 Pg C yr^{-1}, implying a carbon-neutral residual (that is, approximately zero NBE). On the other hand, for 2011 when NBE was -0.25 ± 0.14 Pg C yr^{-1}, the overall carbon balance was neutral, because this was offset by fire-associated losses of roughly the same size (0.30 ± 0.10 Pg C yr^{-1}). The return of

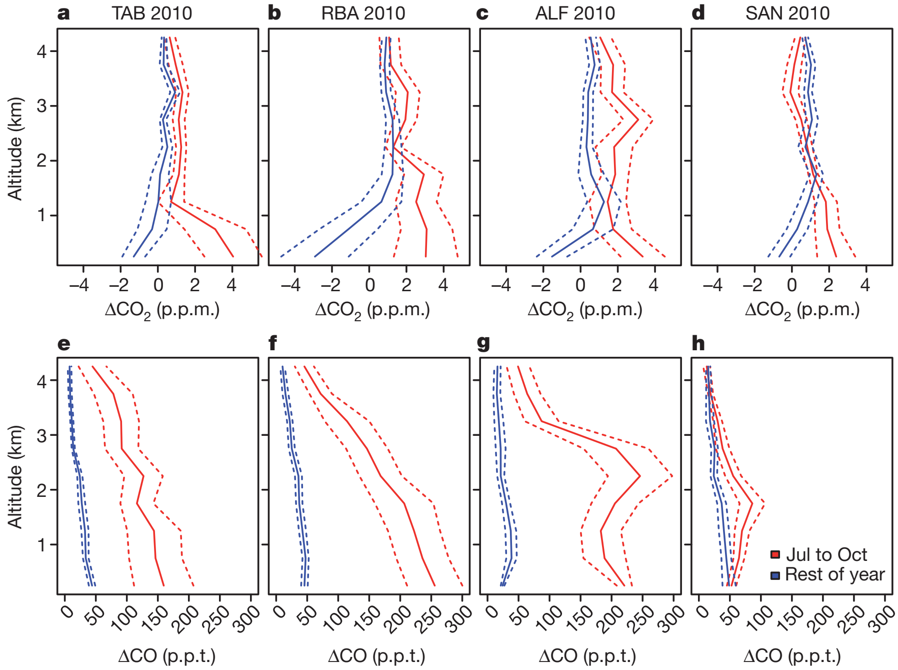


Figure 3 | Surface flux signals in vertical profiles. **a**–**d**, Mean difference between CO_2 profiles measured in 2010 at the four Amazonian aircraft sampling sites and oceanic CO_2 background (that is, ΔCO_2) during the dry (red lines) and wet (blue lines) seasons, respectively (solid lines) and the standard deviation divided by the square root of number of profiles (dashed lines). The background is estimated from in situ SF_6 and CO_2 at the NOAA/ESRL monitoring stations ASC and RPB, as described in the main text. **e**–**h**, As for **a**–**d**, but for CO. p.p.t., parts per trillion. The dry season (red lines) is affected by fires at most sites and is here defined as July–October for illustrative purposes only; it does not correspond to all months with fire emissions (see Methods).

the unburned Amazonian vegetation to being a sink in 2011 seems to have been driven primarily by precipitation, which changed from a negative anomaly in 2010 to a positive anomaly in 2011 (Extended Data Fig. 1a, b). However, temperatures were higher than average for both years, reflecting a net warming trend in recent decades (Extended Data Fig. 1c, d).

A more detailed picture of the Amazonian carbon cycle response to climate is revealed by the quarterly fluxes and by focusing first on RBA, TAB and ALF. For both years, during the first quarter of the year (the start of the wet season), measurements indicate a net carbon sink, and during the second and drier half of the year, measurements indicate a net source (Fig. 4a). However, during the second quarter of 2010 (in

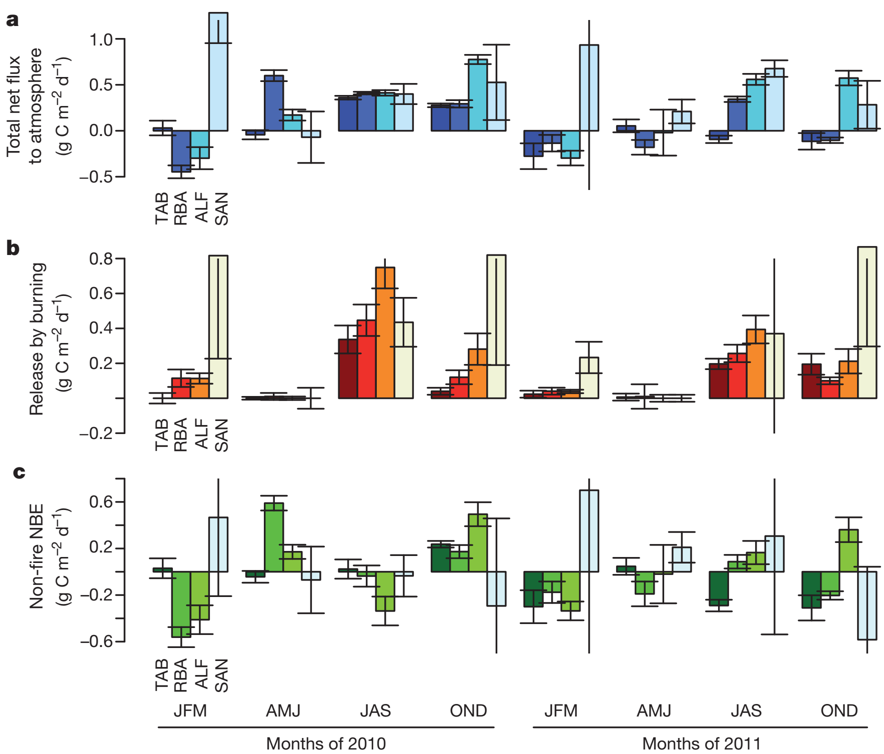


Figure 4 | Flux estimates summary. Quarterly flux and standard error (see Methods) of total carbon flux to the atmosphere (**a**), carbon release due to biomass burning (**b**) and carbon loss from the land (NBE) (**c**) based on the sites TAB, RBA, SAN and ALF for 2010 and 2011.

Table 1 | Summary of annual carbon flux estimates

Sites	TAB	RBA	SAN	ALF	
	2010 fluxes ($g C m^{-2} d^{-1}$)				Scaled 2010 flux ($Pg C yr^{-1}$)†
Total	0.15 ± 0.10	0.17 ± 0.11	0.33 ± 0.50	0.29 ± 0.15	0.48 ± 0.18
Fire	0.13 ± 0.05	0.17 ± 0.06	0.57 ± 0.45	0.28 ± 0.09	0.51 ± 0.12
NBE	0.02 ± 0.11	0.00 ± 0.13	−0.25 ± 0.70	0.01 ± 0.17	−0.03 ± 0.22
	2011 fluxes ($g C m^{-2} d^{-1}$)				Scaled 2011 flux ($Pg C yr^{-1}$)†
Total	−0.10 ± 0.07	−0.04 ± 0.07	0.46 ± 0.20	0.24 ± 0.06	0.06 ± 0.10
Fire	0.08 ± 0.03	0.09 ± 0.03	0.44 ± 0.51	0.16 ± 0.04	0.30 ± 0.10
NBE	−0.18 ± 0.08	−0.13 ± 0.08	0.02 ± 0.84	0.08 ± 0.07	−0.25 ± 0.14
Area of influence ($\times 10^6 km^2$)*	2.53	3.67	0.59	1.31	

The uncertainties are standard errors calculated by propagating uncertainties in all equations using a Monte Carlo approach, and then taking half the value of the 16th–84th percentile range. A bootstrapping approach to calculate the standard error (2.5th–97.5th percentile range) yields slightly smaller values.
* Back-trajectory ensemble envelope (that is, the total area of influence of a measuring site as estimated from wind back-trajectory ensembles).
† 'Scaled' means the flux estimates have been scaled to the tropical South America forested area, assuming an Amazon forest area of $6.77 \times 10^6 km^2$ (ref. 30).

contrast to 2011) we calculate the flux to be a carbon source, which slightly lags the strong precipitation and temperature anomalies in February and March. Net emissions during the second half of 2010 were more than twice as large as in 2011, corresponding to precipitation and temperature anomalies in August and September 2010. For both years, however, the difference in carbon release between the second and first half of the year is mainly due to fire emissions (Fig. 4 and Extended Data Fig. 2). The larger fire emissions in 2010 are consistent with the anomalously high fire counts observed from space (Fig. 2b, Extended Data Figure 2) and basin-wide CO anomalies, which in 2010 extended well above ~2 km a.s.l. (roughly the planetary boundary layer height) into the free troposphere, even at the more remote sites RBA and TAB (Fig. 3e–h and Extended Data Fig. 2). Moreover, the 'arc of deforestation' in the southern and eastern Amazon basin was one of the regions with the strongest precipitation anomalies (Extended Data Fig. 1a, b), intensifying the meteorological conditions required for fire ignition and persistence, and probably leading to the large burning emissions we observed in 2010. After accounting for fire emissions, the residual NBE reveals large differences between the years, especially for the second and fourth quarters, for which there were large carbon releases in 2010 but smaller ones in 2011. This difference in seasonality between the two years appears to reflect a lagged drought stress induced by precipitation anomalies in February/March (first quarter) and August/September (third quarter) of 2010.

The fluxes calculated from the SAN data differ from the other three sites both in seasonality and in the contrast between 2010 and 2011, with a strong carbon source in the first quarter of the year for air sampled upwind of SAN (but not the other three sites) especially notable. This may result in part from the fire season extending into January for the eastern Amazon and northeast Brazil, which is not the case for the moister central/western areas. Additionally, eddy-flux data[11] and CO_2 vertical profile analysis[20] show that (unburned) forests in the eastern Amazon are net sinks in the dry season and net sources in the wet season. In contrast, other sites tend to show wet season uptake (Figs 3a–d and 4).

Additional insight about the cause of the difference in 2010 and 2011 NBE comes from observations at a network of 14 intensive forest carbon cycle measurement plots established across the Amazon basin. At these plots a near-complete suite of carbon pools is being observed, providing an estimate of net primary production and autotrophic respiration and thus an upper bound on gross primary production[5]. Six of these plots experienced anomalous drought stress in 2010, at which time gross primary production declined (Extended Data Fig. 5a), and there were minimal positive temperature anomalies (Extended Data Fig. 5b). Combined, atmospheric mass balance and forest plot analysis suggest that drought has an important negative effect on Amazon forest productivity and with likely consequences on future changes in the forests. This is in contrast to a recent analysis of future Amazon carbon losses calibrated via inter-annual responses of global atmospheric CO_2 growth rates to tropical temperature anomalies[24].

Tropical temperature anomalies have tended to covary with moisture anomalies in the past, so although these models seem to reproduce recent variability correctly they may do so for the wrong reason. Moreover, as 2011 shows, positive temperature anomalies can also coincide with non-drought years.

Besides the new insights into large-scale controls of carbon pool responses in a changing climate, our results provide a top-down confirmation that during non-drought years intact Amazonian forests are a substantial carbon sink, consistent with theoretical predictions for forest biomass alone[25]. Our NBE estimate for 2011 is smaller than the mean annual biomass sink of 0.39 ± 0.10 Pg C estimated for the 1980–2004 period based on repeated censuses at a widespread forest plot network[7]. However, our fire flux estimate is not identical to the total deforestation emissions, which includes emissions from heterotrophic respiration, thus slightly biasing our NBE estimate. The Deforestation Carbon Flux (DECAF) land-use change model[26] suggests that the sources of deforestation emissions in the southern Amazon are typically 30% respiration and 70% fire, implying 2011 deforestation fluxes of about $+0.4 Pg C yr^{-1}$, and therefore NBE of about $-0.4 Pg C yr^{-1}$, closing the gap between the top-down and bottom-up estimates. In 2011 in particular, respiration could have been stimulated following enhanced tree mortality caused by the 2010 drought[27].

In summary, we have empirically documented a pronounced response of a large fraction of the Amazonian vegetation to drought, with forest productivity stalled and large amounts of carbon released by fire in 2010. The Amazon basin returned to being a net carbon sink in 2011. But our results are cause for concern in the light of the recent increase in precipitation extremes and increasing temperatures. If these climate trends continue, future shifts in Amazon forest function, leading to reduced carbon uptake, are likely. This could exacerbate carbon losses as a result of direct human activities such as deforestation.

METHODS SUMMARY

Air sample profiles were taken using small aircraft descending in a spiral from approximately 4,420 m to about 300 m a.s.l. (as close to the forest canopy as possible), semi-automatically filling 12 (for the TAB, ALF and RBA sites) and 17 (for the SAN site) 0.7-litre flasks controlled from a microprocessor and contained in one suitcase. Profiles are taken between 12:00 and 13:00 local time. At that time, the boundary layer is close to being fully developed. Once a vertical profile has been sampled (one suitcase filled) it is transported to the IPEN Atmospheric Chemistry Laboratory in Sao Paulo, where samples are analysed by a replica of the NOAA/ESRL trace gas analysis system. All aircraft data used in this study is available at ftp://ftppub.ipen.br/nature_gatti_etal/. The accuracy and precision of the system are evaluated with three independent procedures that demonstrate excellent performance with long-term repeatability (1σ) of ±0.03 parts per million (p.p.m.) and a difference between measured and calibrated values of 0.03 p.p.m. Because NOAA/ESRL Atlantic data from the ASC and RPB sites are used as background values for Amazonian measurements made at IPEN, this high accuracy is required to ensure that spatial gradients are not artefacts of calibration. The CO and SF_6 measurements presented here are also made at IPEN with calibration standards

tied directly to the World Meteorological Organization reference scales maintained by NOAA/ESRL.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 May; accepted 12 December 2013.

1. Huntingford, C. *et al.* Contributions of carbon cycle uncertainty to future climate projection spread. *Tellus B* **61,** 355–360 (2009).
2. Friedlingstein, P. *et al.* Climate-carbon cycle feedback analysis: results from the (CMIP)-M-4 model intercomparison. *J. Clim.* **19,** 3337–3353 (2006).
3. Phillips, O. L. *et al.* Changes in the carbon balance of tropical forests: evidence from long-term plots. *Science* **282,** 439–442 (1998).
4. Huntingford, C. *et al.* Simulated resilience of tropical rainforests to CO_2-induced climate change. *Nature Geosci.* **6,** 268–273 (2013).
5. Malhi, Y. *et al.* The regional variation of aboveground live biomass in old-growth Amazonian forests. *Glob. Change Biol.* **12,** 1107–1138 (2006).
6. Gloor, M. *et al.* Intensification of the Amazon hydrological cycle over the last two decades. *Geophys. Res. Lett.* **40,** 1729–1733 (2013).
7. Phillips, O. L. *et al.* Drought sensitivity of the Amazon rainforest. *Science* **323,** 1344–1347 (2009).
8. Lloyd, J. *et al.* An airborne regional carbon balance for Central Amazonia. *Biogeosciences* **4,** 759–768 (2007).
9. Chou, W. W. *et al.* Net fluxes of CO_2 in Amazonia derived from aircraft observations. *J. Geophys. Res.* **107,** 4614, http://dx.doi.org/10.1029/2001JD001295 (2002).
10. Saleska, S., da Rocha, H. R. & Nobre, A. in *Amazonia and Global Change Geophysical Monograph Series* **186,** 389–407 (ed. Gash, J., Keller, M. & Silva Dias, P.) (American Geophysical Union, 2009).
11. Saleska, S. R. *et al.* Carbon in Amazon forests: unexpected seasonal fluxes and disturbance-induced losses. *Science* **302,** 1554–1557 (2003).
12. Houghton, R. A. Revised estimates of the annual net flux of carbon to the atmosphere from changes in land use and land management 1850–2000. *Tellus B* **55,** 378–390 (2003).
13. Richey, J. E., Melack, J. M., Aufdenkampe, A. K., Ballester, V. M. & Hess, L. L. Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO_2. *Nature* **416,** 617–620 (2002).
14. Gurney, K. R. *et al.* Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models. *Nature* **415,** 626 (2002).
15. Stephens, B. B. *et al.* Weak northern and strong tropical land carbon uptake from vertical profiles of atmospheric CO_2. *Science* **316,** 1732–1735 (2007).
16. Aragão, L. E. O. C. *et al.* Spatial patterns and fire response of recent Amazonian droughts. *Geophys. Res. Lett.* **34,** L07701, http://dx.doi.org/10.1029/2006gl028946 (2007).
17. Espinoza, J. C. *et al.* Climate variability and extreme drought in the upper Solimões River (western Amazon Basin): understanding the exceptional 2010 drought. *Geophys. Res. Lett.* **38,** L13406, http://dx.doi.org/10.1029/2011gl047862 (2011).
18. Miller, J. B. *et al.* Airborne measurements indicate large methane emissions from the eastern Amazon basin. *Geophys. Res. Lett.* 34, L10809, http://dx.doi.org/10.1029/2006GL029213 (2007).
19. D'Amelio, M. T. S., Gatti, L. V., Miller, J. B. & Tans, P. Regional N_2O fluxes in Amazonia derived from aircraft vertical profiles. *Atmos. Chem. Phys.* **9,** 8785–8797 (2009).
20. Gatti, L. V. *et al.* Vertical profiles of CO_2 above eastern Amazonia suggest a net carbon flux to the atmosphere and balanced biosphere between 2000 and 2009. *Tellus B* **62,** 581–594 (2010).
21. European Commission. Emission Database for Global Atmospheric Research (EDGAR) version 4.0, http://edgar.jrc.ec.europa.eu/overview.php?v=40 (Joint Research Centre/Netherlands Environmental Assessment Agency, 2009).
22. Greenberg, J. P. *et al.* Biogenic VOC emissions from forested Amazonian landscapes. *Glob. Change Biol.* **10,** 651–662 (2004).
23. Gloor, M. *et al.* The carbon balance of South America: a review of the status, decadal trends and main determinants. *Biogeosciences* **9,** 5407–5430 (2012).
24. Cox, P. M, *et al.* Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature* **494,** 341–344 (2013).
25. Lloyd, J. & Farquhar, G. D. The CO_2 dependence of photosynthesis, plant growth responses to elevated atmospheric CO_2 concentrations and their interaction with soil nutrient status. I. General principles and forest ecosystems. *Funct. Ecol.* **10,** 4–32 (1996).
26. van der Werf, G. R. *et al.* Estimates of fire emissions from an active deforestation region in the southern Amazon based on satellite data and biogeochemical modelling. *Biogeosciences* **6,** 235–249 (2009).
27. Lewis, S. L., Brando, P. M., Phillips, O. L., van der Heijden, G. M. F. & Nepstad, D. The 2010 Amazon drought. *Science* **331,** 554 (2011).
28. Liu, Z., Ostrenga, D., Teng, W. & Kempler, S. Tropical Rainfall Measuring Mission (TRMM) precipitation data and services for research and applications. *Bull. Am. Meteorol. Soc.* **93,** 1317–1325 (2012).
29. Arino, O., Casadio, S. & Serpe, D. Global night-time fire season timing and fire count trends using the ATSR instrument series. *Remote Sens. Environ.* **116,** 226–238 (2012).
30. INPE. *PRODES (Projeto de Deflorestamento da Amazônia)* http://www.obt.inpe.br/prodes/index.html (2011).

Acknowledgements We thank P. Tans and P. Bakwin, who had the foresight to initiate a long-term high-precision greenhouse gas measurement laboratory in Sao Paulo, and D. Wickland, the NASA programme manager who initially supported this effort. This work has been financed primarily by the UK Environmental Research Council (NERC) via the consortium grant 'AMAZONICA' NERC (NE/F005806/1) and also by the State of Sao Paulo Science Foundation (FAPESP) via the 'Carbon Tracker' project (08/58120-3), and the EU via the 7th grant framework GEOCARBON project (grant number agreement 283080). NASA, NOAA and IPEN made large contributions to the construction and maintenance of the GHG laboratory in Brazil. Intensive plot measurements were supported by NERC and the Moore Foundation via grants given to RAINFOR. L.G.D., L.S.B., C.S.S.C., V.F.B. and A.M. were supported by CNPq, CAPES, Fapesp and IPEN, and O.L.P. by an ERC Advanced Grant. We thank measurement analysts and scientists at NOAA for providing data, and the pilots who collected the air samples. Numerous people at NOAA, especially A. Crotwell, D. Guenther, C. Sweeney and K. Thoning, provided advice and technical support for air sampling and measurements in Brazil. E. Dlugokencky provided data from Ascension Island and Ragged Point in Barbados. We also thank D. Galbraith for help with the comprehensive forest census plot data and R. Brienen for comments. Finally, we acknowledge S. Denning for reviews of the manuscript.

Author Contributions L.V.G., M.G., J.B.M., J.L., H.R., O.L.P., Y.M. and J.G. conceived the basin-wide measurement programme and approach. M.G., J.B.M. and L.V.G. wrote the paper. C.E.D. and Y.M. analysed and contributed the data of the comprehensive biometric forests census plots. S.F., R.B., L.O.A., L.G.D. and L.S.B. helped with data analysis. V.F.B., C.S.C.C. and A.M. helped with greenhouse gas concentration analysis. All co-authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.V.G. (lvgatti@gmail.com), M.G. (eugloor@gmail.com) and J.B.M. (john.b.miller@noaa.gov).

LETTER

doi:10.1038/nature12924

Convective forcing of mercury and ozone in the Arctic boundary layer induced by leads in sea ice

Christopher W. Moore[1]*, Daniel Obrist[1]*, Alexandra Steffen[2], Ralf M. Staebler[2], Thomas A. Douglas[3], Andreas Richter[4] & Son V. Nghiem[5]

The ongoing regime shift of Arctic sea ice from perennial to seasonal ice is associated with more dynamic patterns of opening and closing sea-ice leads (large transient channels of open water in the ice)[1–3], which may affect atmospheric and biogeochemical cycles in the Arctic[4]. Mercury and ozone are rapidly removed from the atmospheric boundary layer during depletion events in the Arctic[5–7], caused by destruction of ozone along with oxidation of gaseous elemental mercury (Hg(0)) to oxidized mercury (Hg(II)) in the atmosphere and its subsequent deposition to snow and ice[5]. Ozone depletion events can change the oxidative capacity of the air by affecting atmospheric hydroxyl radical chemistry[8], whereas atmospheric mercury depletion events can increase the deposition of mercury to the Arctic[6,9–11], some of which can enter ecosystems during snowmelt[12]. Here we present near-surface measurements of atmospheric mercury and ozone from two Arctic field campaigns near Barrow, Alaska. We find that coastal depletion events are directly linked to sea-ice dynamics. A consolidated ice cover facilitates the depletion of Hg(0) and ozone, but these immediately recover to near-background concentrations in the upwind presence of open sea-ice leads. We attribute the rapid recoveries of Hg(0) and ozone to lead-initiated shallow convection in the stable Arctic boundary layer, which mixes Hg(0) and ozone from undepleted air masses aloft. This convective forcing provides additional Hg(0) to the surface layer at a time of active depletion chemistry, where it is subject to renewed oxidation. Future work will need to establish the degree to which large-scale changes in sea-ice dynamics across the Arctic alter ozone chemistry and mercury deposition in fragile Arctic ecosystems.

Profound changes that have occurred recently in the Arctic sea ice include historic minimum extents of perennial sea ice[1] and a shift to thinner seasonal sea ice[1,2], which experiences more dynamic patterns of opening and closing sea-ice leads[2]. These changes have consequences for the Arctic energy balance and the Earth's radiation budget, with a positive feedback that can accelerate Arctic warming[3]. Here we show that atmospheric mercury (Hg) and ozone (O_3) depletion events near Barrow, Alaska, are directly linked to sea-ice dynamics in the Beaufort and Chukchi seas. We performed near-surface measurements of atmospheric Hg and O_3 directly over the frozen Chukchi Sea during two field studies: the Bromine, Ozone, and Mercury Experiment (BROMEX)[13] in March/April 2012, and the Ocean-Atmosphere-Sea Ice-Snowpack (OASIS) campaign[14] in March 2009 (Fig. 1). We characterized the surrounding sea-ice conditions with daily Moderate Resolution Imaging Spectroradiometer (MODIS) satellite images and marked the location of open leads in the path of air masses during the previous 24 hours before the air masses arrived at the site. We consistently observed that periods of strong and concurrent atmospheric Hg depletion events (below 0.8 ng m^{-3}) and O_3 depletion events (below 5 parts per billion by volume (p.p.b.v.)) occurred when upwind areas consisted largely of consolidated sea-ice cover (completely frozen or containing fully refrozen leads). Periods when air masses travelled over open leads within about 150 km upwind of Barrow, however, were associated with higher, undepleted Hg(0) and O_3 concentrations (Fig. 1).

Using high-temporal-resolution (every 4 hours) National Oceanic and Atmospheric Administration (NOAA) Hybrid Single Particle Lagrangian Integrated Trajectory Model (HYSPLIT) back-trajectories, we show the effects of sea-ice leads on boundary layer Hg(0) and O_3 for several periods associated with dramatic changes in Hg(0) and O_3 concentrations (Figs 2 and 3 and Extended Data Figs 1 and 2). The first period in 2012 (Fig. 2) shows strong increases in Hg(0) and O_3 concentrations when back-trajectories switched from areas dominated by consolidated sea ice to areas with open leads. Initially (cases 1 and 2), back-trajectories travelled entirely over consolidated sea ice; and although open leads occurred north of Barrow, back-trajectories did not intersect with these within the previous 24 hours. Therefore, the open leads were not affecting atmospheric Hg(0) and O_3 concentrations for these two cases. During this period, Hg(0) and O_3 concentrations in the Arctic atmospheric boundary layer were depleted (<0.6 ng m^{-3} and <15 p.p.b.v., respectively), indicating an ongoing atmospheric Hg and O_3 depletion event. A new, 2-km-wide lead opened northeast of Barrow on 24 March 2012 (case 3). Although back-trajectories changed little since the two previous days (also supported by consistent wind velocities, see Extended Data Fig. 3), they now crossed this open lead and concentrations of Hg(0) and O_3 dramatically increased to 1.2 ng m^{-3} and 33 p.p.b.v. within hours, approaching Northern Hemisphere background concentrations (roughly 1.5 ng m^{-3} for Hg(0) and 30 p.p.b.v. for O_3).

In a second example period in 2009 (Fig. 3), open leads were present close to Barrow on 13 March, but air masses initially travelled over areas consisting of consolidated sea ice (case 4). This period was marked by decreasing O_3 concentrations (starting at 40 p.p.b.v. and decreasing to <5 p.p.b.v.), showing an O_3 depletion event. Early on 14 March, O_3 concentrations increased threefold within 1–2 hours, and Hg(0) was correspondingly at near-background concentrations, exactly when the back-trajectories crossed a newly opened, 1-km-wide lead northeast of Barrow (cases 5 and 6). When this lead refroze on 14 March and the trajectories later moved south over consolidated sea ice during the next 2 days, concentrations of Hg(0) and O_3 quickly depleted to near instrument detection limits (case 7). Again, on 16 March, concentrations of O_3 quickly increased when air-mass trajectories crossed a newly developed lead 20 km from Barrow (case 8). As this lead continued to widen (case 9), both Hg(0) and O_3 remained near background levels.

Two more time periods when highly dynamic patterns of Hg(0) and O_3 were directly linked to sea-ice lead dynamics are shown in Extended Data Figs 1 and 2, demonstrating a total of 15 cases of this interaction. Patterns were consistent throughout all periods: when air masses travelled over consolidated sea ice or refrozen leads, Hg(0) and O_3 were depleted or showed decreasing concentrations. When air masses crossed open leads within the 24 hours before measurements, Hg(0) and O_3

[1]Division of Atmospheric Sciences, Desert Research Institute, Reno, Nevada 89523, USA. [2]Air Quality Processes Research Section, Environment Canada, Toronto, Ontario M3H 5T4, Canada. [3]US Army Cold Regions Research and Engineering Laboratory, Fort Wainwright, Alaska 99703, USA. [4]Institute of Environmental Physics, University of Bremen, Bremen 28359, Germany. [5]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA.
*These authors contributed equally to this work.

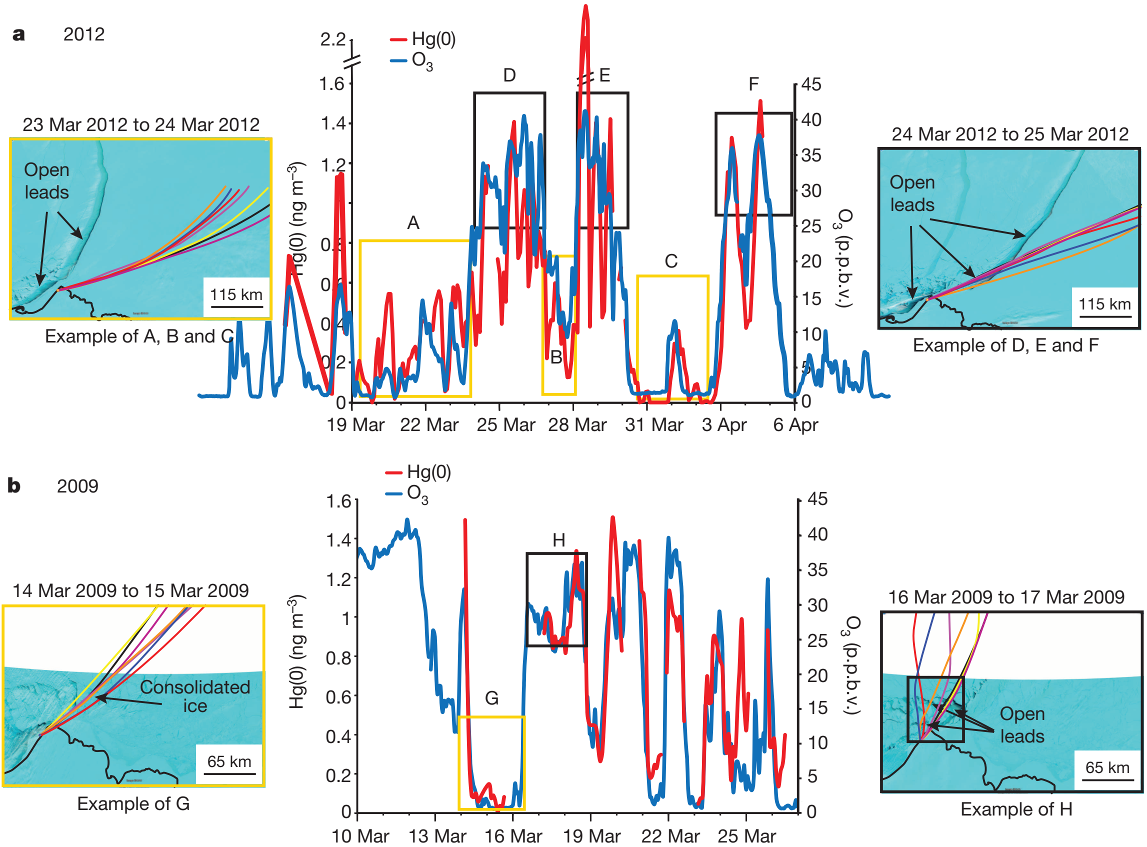


Figure 1 | Time series of Hg(0) and O_3 concentrations. Concentrations of Hg(0) and O_3 for 2012 (**a**) and 2009 (**b**). Yellow boxes are periods when air masses crossed upwind areas of consolidated sea ice. Black boxes are periods when air-mass trajectories crossed open leads. The satellite images represent four typical sea-ice conditions that occurred during measurements. Original satellite images from Google Earth, Terrametrics.

concentrations were not, or were only slightly, depleted. One exception to this pattern occurred during a period after 18 March 2009 (Fig. 1), when the seasonal sea ice surrounding Barrow was characterized by large upwind leads (up to 30 km wide) and showed a complex mixture of frozen surface, open and refrozen leads. Both Hg(0) and O_3 showed dynamic fluctuations between depletion and background levels under these conditions, but the temporal and spatial resolution of the satellite imagery did not allow us to link open leads directly to Hg(0) and O_3 concentrations during that time. Mean concentrations of Hg(0) and O_3 during the eight distinct periods highlighted in Fig. 1 were statistically

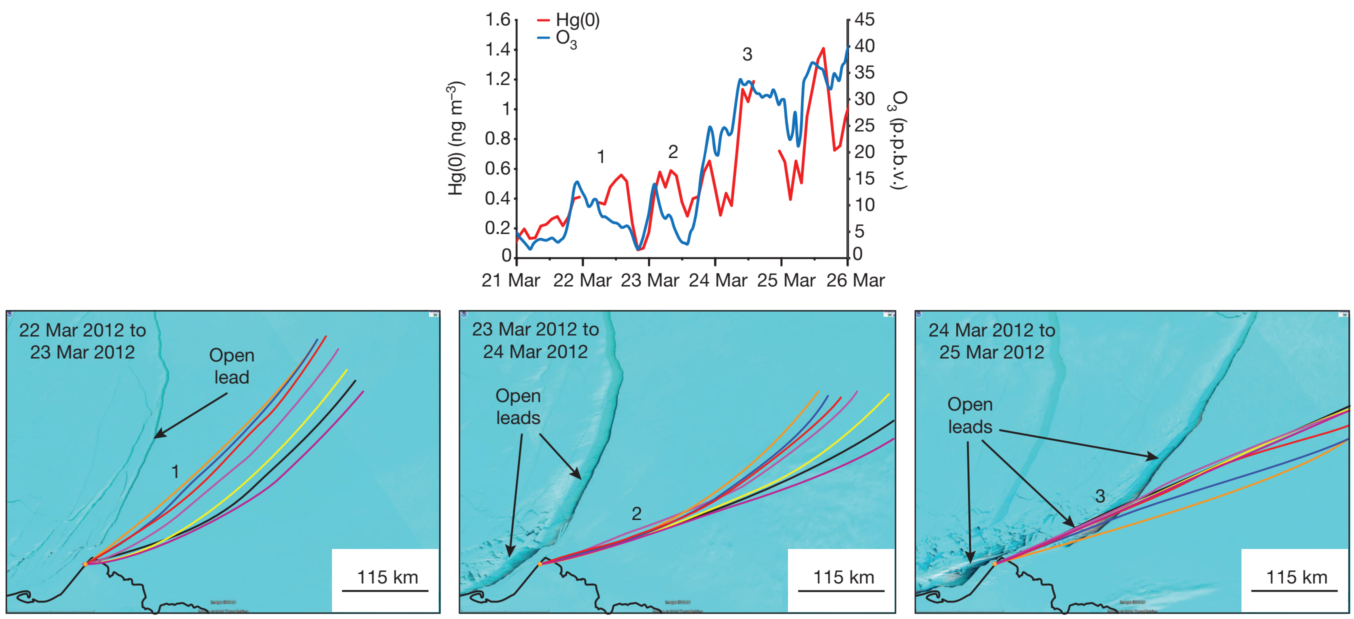


Figure 2 | Impact of sea-ice leads on Hg(0) and O_3 in 2012. Hg(0) and O_3 concentrations between 21 March 2012 and 26 March 2012. Bold numbers correspond to time periods as numbered on the corresponding satellite images. Satellite images were taken at approximately 16:00 UTC (Coordinated Universal Time) each day. Colours represent 24-hour HYSPLIT back-trajectory arrival times near Barrow: orange, 04:00 UTC; blue, 08:00 UTC; red, 12:00 UTC; pink, 16:00 UTC; yellow, 20:00 UTC; black, 00:00 UTC (the next day); and purple, 04:00 UTC (the next day). Original satellite images from Google Earth, Terrametrics.

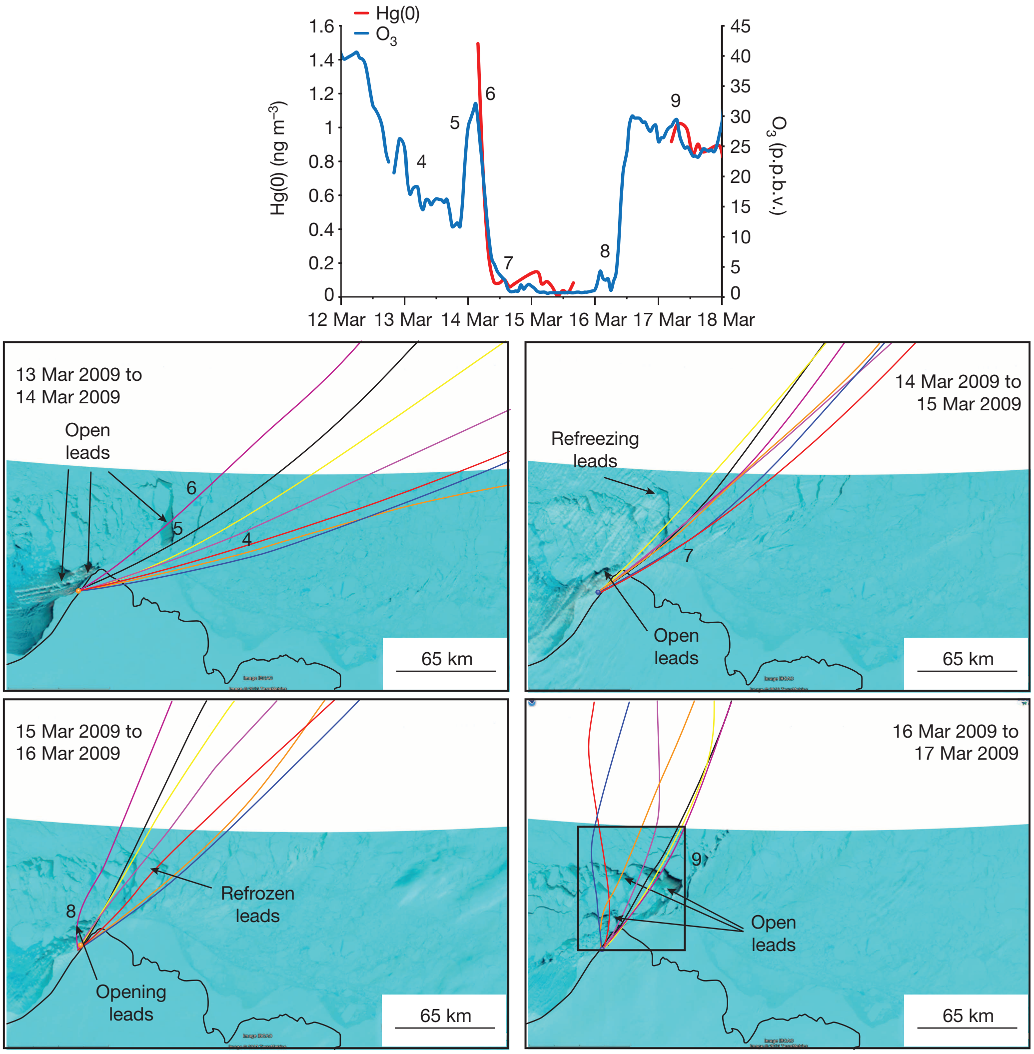


Figure 3 | Impact of sea-ice leads on Hg(0) and O_3 in 2009. Hg(0) and O_3 concentrations between 12 March 2009 and 18 March 2009. Missing Hg(0) concentrations are due to the analyser being moved between locations. Bold numbers correspond to time periods as numbered on the corresponding satellite images. Satellite images were taken at approximately 16:00 UTC each day. Colours represent 24-hour HYSPLIT back-trajectory arrival times near Barrow: orange, 04:00 UTC; blue, 08:00 UTC; red, 12:00 UTC; pink, 16:00 UTC; yellow, 20:00 UTC; black, 00:00 UTC (the next day); and purple, 04:00 UTC (the next day). Original satellite images from Google Earth, Terrametrics.

($P < 0.01$) significantly lower during periods when air masses were unaffected by open leads than during upwind crossing of open leads (Extended Data Table 1).

We considered both chemical and physical processes to explain observed linkages between lead dynamics and boundary-layer chemistry. Measurable levels of gaseous bromine species were reported during both years of measurements[14,15], consistent with active bromine chemistry[16], which is associated with O_3 and atmospheric Hg depletion events. Depletions of Hg(0) and O_3 occurring over consolidated sea ice were probably induced by active bromine chemistry in the area[15], although this process is not yet fully understood[5].

However, we cannot attribute the recoveries of concentrations within 1–2 hours of lead openings to chemical processes only. For example, bromine monoxide (BrO) concentrations, as identified from the Global Ozone Monitoring Experiment-2 (GOME-2)[17] spectrometer, showed a consistent presence of large BrO clouds in the region. Therefore, the patterns of Hg(0) and O_3 were unrelated to coincidental patterns at the edges of BrO clouds (Extended Data Fig. 4). This was supported by direct measurements of BrO concentrations at the site that were not significantly correlated with Hg(0) or O_3. Even if atmospheric Hg- and O_3-depletion-event chemistry were to stop on contact with open leads, the Hg(0) and O_3 concentrations would remain depleted for some time and would not quickly recover within just a few hours. Sources of Hg leading to partial recovery of Hg(0) after atmospheric Hg depletion events could include the photochemical reduction of Hg(II) and re-emission from surfaces[5,18,19], but no such source exists for O_3 because O_3 is destroyed during oxidation. Another possible source for Hg(0) recovery involves emissions from Arctic Ocean water[20,21]. However, this would not explain the simultaneous recovery of both Hg(0) and O_3 because O_3 is not typically emitted by the ocean. It is also striking that recoveries of both Hg(0) and O_3 consistently reached levels near Northern Hemisphere background concentrations, independently of the size of the leads. If there were an ocean source of Hg, this would not be expected.

We attribute the fast transitions from depleted to non-depleted Hg(0) and O_3 levels to changes in boundary-layer dynamics induced by sea-ice leads, which dominate the effects of underlying depletion chemistry. Lead openings generate large sensible and latent heat fluxes from the water surface to the atmosphere owing to strong temperature gradients (more than 20 K) between the warmer ocean water and cold polar atmosphere[22,23]. This heat transfer causes significant convective mixing in the atmosphere directly above and downwind of leads[23,24] (see video in ref. 13 of the lead cloud recorded during BROMEX). We propose that such convective mixing produces fast recoveries of surface Hg(0) and O_3 from air masses aloft. Vertical measurements of Hg(0) and O_3 in the stable polar boundary layer have shown that Hg(0) and O_3 depletions are limited to the surface layer, whereas air aloft is not depleted[7,25,26]. We confirmed increased turbulent mixing using radiosonde data from Barrow during several periods when shallow boundary layers quickly grew in height in the presence of open sea water (for example, see Extended Data Fig. 3). It is also unlikely that increased wind speed alone—often associated with, and a cause of, opening sea-ice leads—would explain Hg(0) and O_3 recoveries through increased wind shear given the periods of Hg(0) and O_3 recoveries when wind speeds changed little and remained low (below 3 m s^{-1}; Extended Data Fig. 5).

The implications of the observed effects of the dynamics of sea-ice leads on atmospheric Hg and O_3 are large: the recovery of Hg(0) and O_3 via convective transport of Hg(0) and O_3 induced by open leads is probably a source of additional Hg(0) and O_3 to the atmospheric surface layer in the Arctic, all other factors remaining unchanged. Once in the surface layer, resupplied O_3 and Hg(0) from aloft can participate in renewed depletion chemistry, as the sea-ice leads occur at a time of active depletion chemistry, possibly increasing deposition loads attributed to Hg depletion events[9,11] and the total amount of O_3 destroyed in the atmosphere. It is possible that a warming environment and changes in sea-ice cover produce other changes in Arctic chemistry that may ultimately affect the quantity of Hg accumulating in biota (for example, shorter duration of sea-ice cover may cause increased photochemical reduction and photo-degradation of methyl mercury[27]). As seasonal sea ice increases at the expense of perennial sea ice and lead activity is expected to increase[1], large areas across the Arctic may experience increased convective replenishment of Hg(0) and O_3, affecting Hg oxidation and O_3 depletion chemistry, as observed in Barrow. Lead-induced shallow convective mixing of air in response to sea-ice leads—as shown for Hg(0) and O_3—could also affect the boundary-layer input of other pollutants—such as persistent organics, aerosols and other heavy metals[28–30].

METHODS SUMMARY

Ground measurements. Ground-based measurements during BROMEX included characterization of speciated atmospheric Hg (including Hg(0), $Hg(II)_{gaseous}$, and $Hg(II)_{particulate}$) and O_3 on the frozen Chukchi Sea (2 km off the coast) and on the frozen tundra (5 km inland), but only data from the sea ice were used for this study. During the OASIS campaign, atmospheric Hg speciation was measured at three different locations over the frozen Arctic Ocean[14]. All the data presented passed strict quality assurance and control protocols, and the Hg(0) concentrations presented are hourly averages. To be consistent, all O_3 concentrations used for both years were from the NOAA-operated Barrow Observatory. Correlation analysis showed excellent agreement of data measured at all stations.

Satellite images and HYSPLIT modelling. Daily, densely gridded (250 m) MODIS images of ice conditions were composed of the 7-2-1 bands (2,105–2,155 nm, 841–876 nm and 620–670 nm wavelengths) from the NASA Terra satellite. These were combined with high-temporal-resolution NOAA HYSPLIT air-mass trajectories modelled 24 hours backwards in time and generated every 4 hours based on meteorology data from the Global Data Assimilation System (GDAS) interpolated from a 1.0° by 1.0° grid in latitude and longitude. HYSPLIT model runs were performed at 25 m, 225 m and 400 m. All back-trajectories presented are at 25 m and, owing to atmospheric stability, were verified at the heights of 225 m and 400 m. HYSPLIT trajectories generated with GDAS meteorological data were also checked with back-trajectories generated from Weather Research and Forecasting Model meteorological data to verify their paths. Daily satellite images were used for both years to map sea-ice conditions precisely over several hundred kilometres around our measurement domain—including solid sea ice, open sea-ice leads and refreezing of previously open leads. The images were overlaid with the HYSPLIT back-trajectories to assess how O_3 depletion events and atmospheric Hg depletion events measured on the ground near Barrow related to the upwind footprint area of measured air masses.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 June; accepted 29 November 2013.
Published online 15 January 2014.

1. Nghiem, S. V. *et al.* Rapid reduction of Arctic perennial sea ice. *Geophys. Res. Lett.* **34,** L19504 (2007).
2. Hutchings, J. K. & Rigor, I. G. Role of ice dynamics in anomalous ice conditions in the Beaufort Sea during 2006 and 2007. *J. Geophys. Res.* **117,** C00E04 (2012).
3. Stroeve, J. *et al.* The Arctic's rapidly shrinking sea ice cover: a research synthesis. *Clim. Change* **110,** 1005–1027 (2012).
4. Nghiem, S. V. *et al.* Field and satellite observations of the formation and distribution of Arctic atmospheric bromine above a rejuvenated sea ice cover. *J. Geophys. Res.* **117,** D00S05 (2012).
5. Steffen, A. *et al.* A synthesis of atmospheric mercury depletion event chemistry in the atmosphere and snow. *Atmos. Chem. Phys.* **8,** 1445–1482 (2008).
6. Brooks, S. B. *et al.* The mass balance of mercury in the springtime arctic environment. *Geophys. Res. Lett.* **33,** L13812 (2006).
7. Bottenheim, J. W., Fuentes, J. D., Tarasick, D. W. & Anlauf, K. G. Ozone in the Arctic lower troposphere during winter and spring 2000 (ALERT2000). *Atmos. Environ.* **36,** 2535–2544 (2002).
8. Simpson, W. R. *et al.* Halogens and their role in polar boundary-layer ozone depletion. *Atmos. Chem. Phys.* **7,** 4375–4418 (2007).
9. Skov, H. *et al.* Fate of elemental mercury in the Arctic during atmospheric mercury depletion episodes and the load of atmospheric mercury to the Arctic. *Environ. Sci. Technol.* **38,** 2373–2382 (2004).
10. Drevnick, P. E., Yang, H., Lamborg, C. H. & Rose, N. L. Net atmospheric mercury deposition to Svalbard: estimates from lacustrine sediments. *Atmos. Environ.* **59,** 509–513 (2012).
11. Ariya, P. A, *et al.* The Arctic: a sink for mercury. *Tellus Ser. B* **56,** 397–403 (2004).
12. Durnford, D. & Dastoor, A. The behavior of mercury in the cryosphere: a review of what we know from observations. *J. Geophys. Res.* **116,** D06305 (2011).
13. Nghiem, S. V. *et al.* Studying bromine, ozone, and mercury chemistry in the Arctic. *Eos* **94,** 289–291 (2013).
14. Steffen, A. *et al.* Atmospheric mercury over sea ice during the OASIS-2009 campaign. *Atmos. Chem. Phys.* **13,** 7007–7021 (2013).
15. Pratt, K. A. *et al.* Photochemical production of molecular bromine in Arctic surface snowpacks. *Nature Geosci.* **6,** 351–356 (2013).
16. Stephens, C. R. *et al.* The relative importance of chlorine and bromine radicals in the oxidation of atmospheric mercury at Barrow, Alaska. *J. Geophys. Res.* **117,** D00R11 (2012).
17. Begoin, M. *et al.* Satellite observations of long range transport of a large BrO plume in the Arctic. *Atmos. Chem. Phys.* **10,** 6515–6526 (2010).
18. Dommergue, A., Ferrari, C. P., Poissant, L., Gauchard, P. A. & Boutron, C. F. Diurnal cycles of gaseous mercury within the snowpack at Kuujjuarapik/Whapmagoostui, Quebec, Canada. *Environ. Sci. Technol.* **37,** 3289–3297 (2003).
19. Lindberg, S. E. *et al.* Dynamic oxidation of gaseous mercury in the Arctic troposphere at polar sunrise. *Environ. Sci. Technol.* **36,** 1245–1256 (2002).
20. Aspmo, K. *et al.* Mercury in the atmosphere, snow and melt water ponds in the North Atlantic Ocean during Arctic summer. *Environ. Sci. Technol.* **40,** 4083–4089 (2006).
21. Andersson, M. E., Sommar, J., Gardfeldt, K. & Lindqvist, O. Enhanced concentrations of dissolved gaseous mercury in the surface waters of the Arctic Ocean. *Mar. Chem.* **110,** 190–194 (2008).
22. Andreas, E. L. & Cash, B. A. Convective heat transfer over wintertime leads and polynyas. *J. Geophys. Res.* **104,** 25721–25734 (1999).
23. Marcq, S. & Weiss, J. Influence of sea ice lead-width distribution on turbulent heat transfer between the ocean and the atmosphere. *Cryosphere* **6,** 143–156 (2012).
24. Serreze, M. C. *et al.* Theoretical heights of buoyant convection above open leads in the winter Arctic pack ice cover. *J. Geophys. Res.* **97,** 9411–9422 (1992).
25. Banic, C. M. *et al.* Vertical distribution of gaseous elemental mercury in Canada. *J. Geophys. Res. D* **108,** 4264 (2003).
26. Seabrook, J. A. *et al.* LIDAR measurements of Arctic boundary layer ozone depletion events over the frozen Arctic Ocean. *J. Geophys. Res. D* **116,** D00S02 (2011).
27. Lehnherr, I., St. Louis, V. L., Hintelmann, H. & Kirk, J. L. Methylation of inorganic mercury in polar marine waters. *Nature Geosci.* **4,** 298–302 (2011).
28. Ubl, S., Scheringer, M., Stohl, A., Burkhart, J. F. & Hungerbuhler, K. Primary source regions of polychlorinated biphenyls (PCBs) measured in the Arctic. *Atmos. Environ.* **62,** 391–399 (2012).
29. Shaw, P. M., Russell, L. M., Jefferson, A. & Quinn, P. K. Arctic organic aerosol measurements show particles from mixed combustion in spring haze and from frost flowers in winter. *Geophys. Res. Lett.* **37,** L10803 (2010).
30. Zheng, J., Shotyk, W., Krachler, M. & Fisher, D. A. A 15,800-year record of atmospheric lead deposition on the Devon Island Ice Cap, Nunavut, Canada: natural and anthropogenic enrichments, isotopic composition, and predominant sources. *Glob. Biogeochem. Cycles* **21,** GB2027 (2007).

Acknowledgements This research was supported in part by the National Aeronautics and Space Administration (NASA) Cryospheric Sciences Program (CSP) and by the Desert Research Institute. The Science and Technology Branch of Environment Canada helped fund Hg measurements in 2012 along with the Canadian International Polar Year government programme in 2009. The research at the Jet Propulsion Laboratory, California Institute of Technology, was supported by NASA CSP. We thank Umiaq for field logistic assistance, the Barrow whaling community for beneficial interactions, and the National Oceanic and Atmosphere Administration (NOAA), Global Monitoring Division for the Barrow Observatory data. We gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for provision of the HYSPLIT transport and dispersion model and READY Website (http://www.ready.noaa.gov) used in this publication. We thank K. Pratt and R. Kreidberg for feedback on the manuscript, B. Hatchett and T. Malamakal for help with radiosonde and WRF data, D. Hall and J. Schmaltz for MODIS imagery support, and J. Deary for outstanding field technical support.

Author Contributions C.W.M. and D.O. performed data analysis, participated in the 2012 field campaign, and prepared an initial version of the manuscript. A.S., R.M.S. and T.A.D. were instrumental in both the 2009 and 2012 field campaigns along with data analysis and interpretation. A.R. was responsible for satellite BrO column data retrieval and interpretation. S.V.N. was responsible for satellite data retrieval and interpretation for 2009 and 2012 and participated in the 2012 field campaign. All authors participated in the writing and editing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.W.M. (chris.moore@dri.edu).

doi:10.1038/nature12911

Pathogens and insect herbivores drive rainforest plant diversity and composition

Robert Bagchi[1,2], Rachel E. Gallery[1,3], Sofia Gripenberg[1,4], Sarah J. Gurr[5,6], Lakshmi Narayan[1], Claire E. Addis[1], Robert P. Freckleton[7] & Owen T. Lewis[1]

Tropical forests are important reservoirs of biodiversity[1], but the processes that maintain this diversity remain poorly understood[2]. The Janzen–Connell hypothesis[3,4] suggests that specialized natural enemies such as insect herbivores and fungal pathogens maintain high diversity by elevating mortality when plant species occur at high density (negative density dependence; NDD). NDD has been detected widely in tropical forests[5–9], but the prediction that NDD caused by insects and pathogens has a community-wide role in maintaining tropical plant diversity remains untested. We show experimentally that changes in plant diversity and species composition are caused by fungal pathogens and insect herbivores. Effective plant species richness increased across the seed-to-seedling transition, corresponding to large changes in species composition[5]. Treating seeds and young seedlings with fungicides significantly reduced the diversity of the seedling assemblage, consistent with the Janzen–Connell hypothesis. Although suppressing insect herbivores using insecticides did not alter species diversity, it greatly increased seedling recruitment and caused a marked shift in seedling species composition. Overall, seedling recruitment was significantly reduced at high conspecific seed densities and this NDD was greatest for the species that were most abundant as seeds. Suppressing fungi reduced the negative effects of density on recruitment, confirming that the diversity-enhancing effect of fungi is mediated by NDD. Our study provides an overall test of the Janzen–Connell hypothesis and demonstrates the crucial role that insects and pathogens have both in structuring tropical plant communities and in maintaining their remarkable diversity.

Understanding the mechanisms that allow species to coexist in natural ecosystems is one of the most enduring questions in community ecology. The key challenge is to identify how competitive exclusion is prevented, particularly in situations where large numbers of species share similar resource requirements[10]. This question has special relevance to tropical forest plant communities, which can have exceptional species richness[2,9]. The rapid degradation and destruction of tropical forests[11,12] and the large impact this may have on global biodiversity[1], carbon and water cycling and climate feedbacks[13] makes understanding the mechanisms maintaining and structuring their diversity imperative.

There is compelling evidence that natural enemies, including insect herbivores and fungal and oomycete pathogens (hereafter referred to collectively as pathogens), regulate many plant populations in the tropics[7,14,15] and elsewhere[16,17]. Transmission of natural enemies is more effective between plants growing in areas of high conspecific density, reducing plant survival. The Janzen–Connell hypothesis suggests that this negative density dependence (NDD) will promote plant community diversity by preventing dominant species from competitively excluding other species[3,4]. This hypothesis is one of the most widely invoked explanations for species coexistence, and ultimately high diversity, in plant communities.

Although numerous studies have revealed NDD in plant communities[6,8,18], there is considerably less empirical support for the contention that this will translate into enhanced community diversity[2,9]. A key study[5] in Panama documented NDD at the seed-to-seedling transition in a suite of 53 species, and linked this NDD to increased community diversity. However, the causes of this NDD were not identified. Although reduced herbivory by vertebrates can alter the composition of tropical plant communities[19–21], such effects rarely show NDD, and insect herbivores and pathogens are widely regarded as the most likely causes of NDD leading to enhanced plant diversity[2,9,22]. Despite this, studies demonstrating a causal link between insect- and pathogen-mediated NDD and plant community diversity are lacking.

We compared community diversity of seeds and recruiting seedlings in a tropical forest in Belize, Central America, and investigated whether experimentally excluding natural enemies decreased plant diversity, as predicted by the Janzen–Connell hypothesis. The effective number of species (inverse Simpson's dominance index, $1/D$) among seedlings recruiting in unmanipulated (control) plots was significantly higher than among seeds falling in adjacent seedfall traps ($\Delta\log(1/D) = 0.69 \pm$ s.e.m. $= 0.058$, $t_{107} = 11.91$, $P < 0.001$), corresponding to a doubling of the effective number of plant species at the seed-to-seedling transition. To determine whether insect herbivores or pathogens could be contributing to this increase in diversity we compared the diversity of seedlings growing in control plots (sprayed weekly with water) to plots where we suppressed either insects by spraying an insecticide (Engeo), or pathogens by spraying one of two fungicides, Amistar or Ridomil. Each of the pesticide treatments reduced species diversity, but the effects were only statistically significant for the fungicide Amistar (Fig. 1a; $t_{105} = -2.45$, $P = 0.016$), which reduced the effective number of species by approximately 16%. This result clearly implicates pathogenic fungi in promoting seedling diversity.

Two other changes in the plant community at the seed-to-seedling transition were evident: a shift in species abundances, and altered species composition. These trends were also affected by pesticide treatments. Insecticide treatment increased the total number of recruiting seedlings by a factor of 2.7 compared to the control (Fig. 1b; $t_{105} = -7.67$, $P < 0.001$), demonstrating that plant-feeding insects are a major cause of mortality at this life stage. Although Amistar enhanced seedling recruitment, this effect was marginally nonsignificant ($t_{105} = 1.81$, $P = 0.074$). Dissimilarity in species composition between the seeds and seedlings, measured using the abundance-weighted Morisita–Horn index (R_h)[23], was approximately 87% in the control plots (Fig. 1c). Treating seedlings with insecticide dramatically and significantly reduced this dissimilarity ($t_{105} = -7.86$, $P < 0.001$). The fungicides (Amistar and Ridomil) did not reduce the dissimilarity to seeds significantly, but nevertheless the dissimilarity between the species compositions of the fungicide-treated plots and the control plots was about 20% (Extended Data Fig. 1). Overall, our results suggest that insects disproportionately kill certain plant species, reducing their abundances during the transition from seeds to seedlings. Insects thus strongly influence the structure of plant communities

[1]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. [2]Ecosystem Management Group, Institute of Terrestrial Ecosystems, ETH Zürich, Universitätstrasse 16, 8092 Zürich, Switzerland. [3]School of Natural Resources and the Environment, University of Arizona, Tucson, Arizona 85721, USA. [4]Section of Biodiversity and Environmental Research, Department of Biology, University of Turku, 20014 Turku, Finland. [5]Department of BioSciences, Geoffrey Pope Building, University of Exeter, Exeter EX4 4QD, UK. [6]Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK. [7]Department of Animal and Plant Science, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.

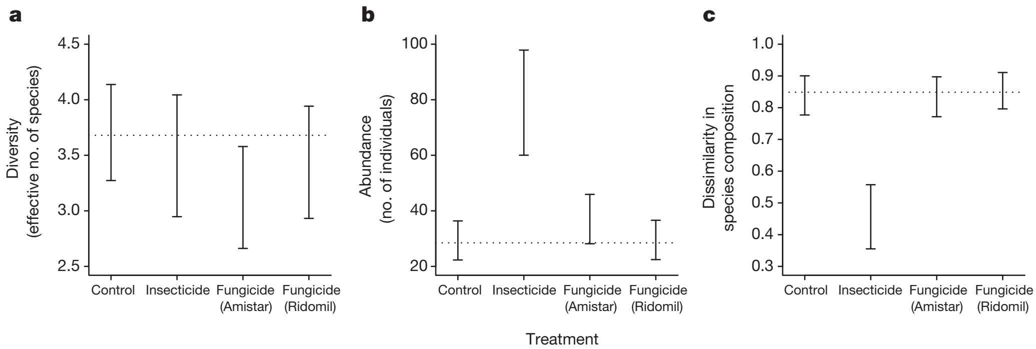


Figure 1 | Suppression of insects and pathogens alters seedling community composition and diversity, respectively. **a**–**c**, Effects of insecticide (Engeo) and two fungicides (Amistar and Ridomil) on the mean effective number of species recruiting as seedlings (**a**); the mean seedling abundance summed across all species (**b**); and the mean abundance-weighted Morisita–Horn dissimilarity in species composition for seedlings under each treatment compared to seeds in adjacent seed traps (**c**). The error bars represent the 95% confidence intervals of the mean across the 36 stations.

in this forest; however, by doing so relatively independently of plant density, their net effect on plant species diversity is small.

For 18 species, sufficient data were available to conduct a formal test for NDD (see Methods). The slope of the relationship between the log number of seeds and the log number of seedlings in the control plots was less than 1 in 13 of the 18 species, and significantly less than 1 for 3 of these (Fig. 2a and Supplementary Table 1), indicating NDD[5]. Furthermore, the mean slope across species was significantly less than 1 in the control treatment ($t_{48} = -2.68$, $P = 0.010$), suggesting that NDD is widespread, as has been found in previous studies of the seed-to-seedling transition in tropical forests[5]. Suppressing fungi using the fungicide Amistar reduced the strength of NDD so that the mean slope was no longer significantly different from 1 (Fig. 2b; $t_{48} = -1.54$, $P = 0.130$). Neither Ridomil nor Engeo reduced the strength of NDD, with the mean slope remaining significantly less than 1 in both treatments (Fig. 2b). Thus, the significant effects of fungal pathogen exclusion on seedling diversity shown in Fig. 1 can be causally linked to a reduction in the magnitude of NDD.

The strength of NDD in control plots was greatest in the species that were most abundant as seeds (Fig. 3a; $t_{16} = -4.33$, $P = 0.001$; Extended Data Table 1). Greater NDD might be detected in these species because their high densities facilitate transmission of insects and pathogens, or because pests and diseases adapt to exploit the most abundant resources. The positive relationship between NDD and seed abundances contrasts with the results of two recent studies investigating NDD in relation to adult abundances[15,24]. These differing results may arise because the rank abundances of species can shift substantially between seeds and adults, and because seed abundance reflects fecundity (which will be inversely correlated with seed size) as well as adult abundance. A third possibility is that abundant seed production is correlated with other traits (for example, lower defence investment or shade tolerance[25]), which are associated in turn with greater susceptibility to density-responsive natural enemies. Regardless of the cause, by reducing the survival of common species disproportionately, NDD may have increased the diversity of recruits more than expected from the average NDD effect. All pesticide treatments weakened the relationship between NDD and abundance markedly (Fig. 3b–d and Extended Data Table 1). By weakening NDD, especially in species that are abundant as seeds, fungicide application may have removed one mechanism for enhancing diversity at the seed-to-seedling transition, leading to the significantly lower seedling diversity observed in the Amistar treatment.

As a final evaluation of the contribution of NDD to enhancing the diversity of recruiting seedlings, we used a simulation approach. Changes in community diversity and composition across the seed-to-seedling transition and following the exclusion of natural enemies could result from either NDD or trade-offs between seed production and allocation to defence against insects and pathogens. To distinguish these two possibilities we used models fitted to the 18 most abundant species to simulate communities in two scenarios (see Methods). In the first scenario (low-density survival), per-capita recruitment was independent of seed density and equal to that expected under each treatment in the absence of conspecific neighbours. In the second scenario (NDD survival), per-capita recruitment was dependent on both pesticide treatment and seed density. Simulations in the low-density survival scenario greatly underestimated the effective number of species in the control plots (Fig. 4a). Total seedling abundance was also overestimated (Fig. 4b) and dissimilarity in species composition underestimated (Fig. 4c) in the control and insecticide treatments in the

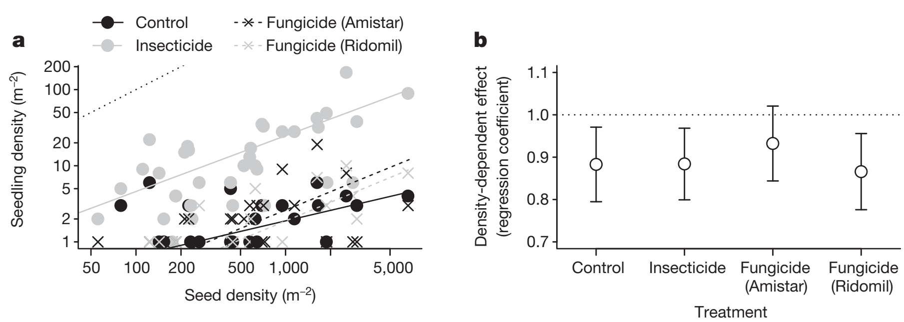


Figure 2 | Recruitment across the seed-to-seedling transition showed NDD in the control, but spraying with the fungicide Amistar removed this NDD. **a**, The relationship between number of recruits and number of seeds for one example species, *Terminalia amazonia*. Without NDD, the expected slope is 1 on a log–log scale (dotted line) and the *y* intercept represents recruitment at low density. The observed slope was lowest (and <1) in the control; treatment with fungicides but not insecticides increased the slope. **b**, The NDD effect is significantly <1 across 18 species in the control treatment, indicating prevalent NDD. Spraying with Amistar, but not other pesticides, removed this effect. Error bars show 95% confidence intervals of the mean.

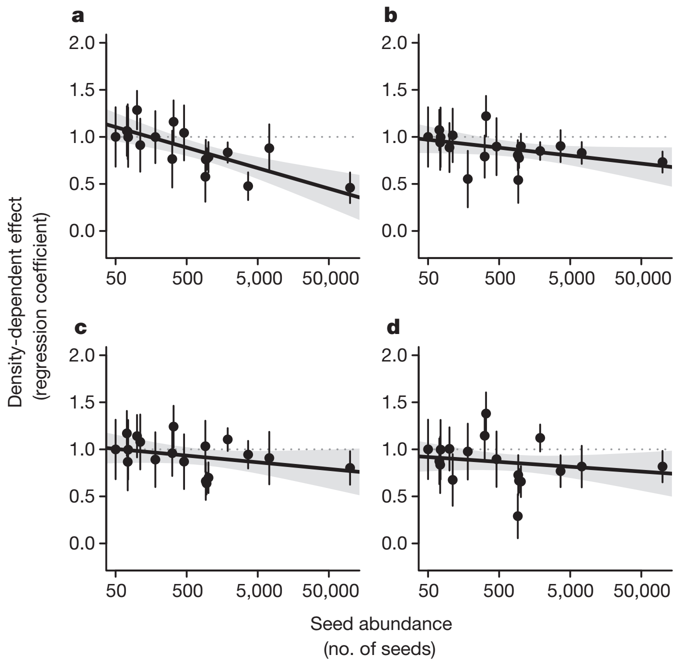


Figure 3 | Negative density dependence is strongest in species that are most abundant as seeds. The relationship between seed abundance and the strength of NDD is shown for the 18 species analysed. **a**–**d**, The relationship for the control plots sprayed with water (**a**), plots sprayed with the insecticide Engeo (**b**), plots sprayed with the fungicide Amistar (**c**), and plots sprayed with an alternative fungicide, Ridomil (**d**). The bold lines are the relationships fitted with a weighted linear mixed-effects model (weights are the inverse of the standard deviations, which are indicated by error bars), with the 95% confidence intervals of the mean. The dotted line shows the null expectation of regression coefficients of 1 in the absence of an effect of seed abundance on the strength of NDD.

absence of NDD. Similar results were obtained using an alternative simulation scenario, where per-capita recruitment reflected that recorded at the mean seed density for each species[5] (Extended Data Fig. 2). Adding NDD to the simulations replicated the observed data better. Overall, these simulations confirm that pathogen-mediated NDD is responsible for increasing the diversity of seedlings.

Although individual components of the Janzen–Connell hypothesis have been tested repeatedly since its formulation more than 40 years ago[3,4], experimental tests of the key overall hypothesis that natural enemies cause NDD and thus promote species coexistence and enhance species diversity are rare. One such study[19] found no evidence that excluding vertebrate herbivores reduced NDD or diversity. However, although vertebrates have occasionally been implicated as drivers of NDD[26–28], the primary causes of NDD are thought to be insects and pathogens[2,9,22]. The results presented here build on existing evidence for widespread NDD in tropical plant communities[5,6,8,18,24] by establishing the cause of NDD, and by linking it to increased plant species diversity, as suggested by the Janzen–Connell hypothesis.

Our experiments highlight that both insect herbivores and pathogens help structure tropical plant communities at the early stages of community assembly and provide support for a pivotal role for natural-enemy-mediated NDD in maintaining species diversity in this tropical forest. Although the magnitude of the NDD we observed was relatively small, this study was conducted over a relatively short timescale (17 months) in a tropical forest of relatively low plant species diversity (approximately 320 tree species have been recorded in the reserve[29]). The effects of NDD will probably accumulate over time, and may be stronger in more species-rich forests. Indeed, similar experiments in other forests are now needed to evaluate the generality of the Janzen–Connell hypothesis as an explanation for variation in species diversity among tropical plant communities.

METHODS SUMMARY

We established 36 sampling stations within a 1-hectare (ha) area in the Chiquibul Forest Reserve, Belize. Each station had three seed traps and four seedling plots. Plots were randomly assigned to four treatments: control (sprayed with water), insecticide (Engeo), or one of two fungicides (Amistar or Ridomil), applied weekly for 17 months. We recorded numbers of seeds from each species collected weekly in each trap. Number and identities of seedlings germinating in each plot were recorded monthly during the peak recruitment period (April to August) and every 2 to 4 months otherwise. We compared the total number of individuals and their diversity (inverse Simpson's dominance index, $1/D$) between seeds and seedlings in the control plot and among pesticide treatments using mixed-effects models. We also compared dissimilarity in species composition between seeds and seedlings (abundance-weighted Morisita–Horn dissimilarity index) among pesticide treatments. In the absence of NDD, a slope of 1 is expected for the relationship between log number of seeds and log number of seedlings[5]. We estimated this slope and the effect of pesticide treatments for 18 species. To determine the average effect of density and the effect of overall species abundance, we modelled the slopes (see Methods) of each species and pesticide treatment combination as a function of log total seed abundance and pesticide treatment, using a weighted mixed-effects model. Finally, to determine whether NDD could generate observed differences in communities among treatments, we used the models of the 18 species to simulate communities, assuming survival to be either density dependent or density independent, based on the establishment probability expected in the absence of conspecific neighbours. We calculated abundance, diversity and dissimilarity based on 1,000 simulations for each scenario and compared the mean and 95% confidence intervals of the observed data to those derived from the simulations. Data (Supplementary Data 1) and code for analyses (Supplementary Notes 1) are provided as Supplementary Information.

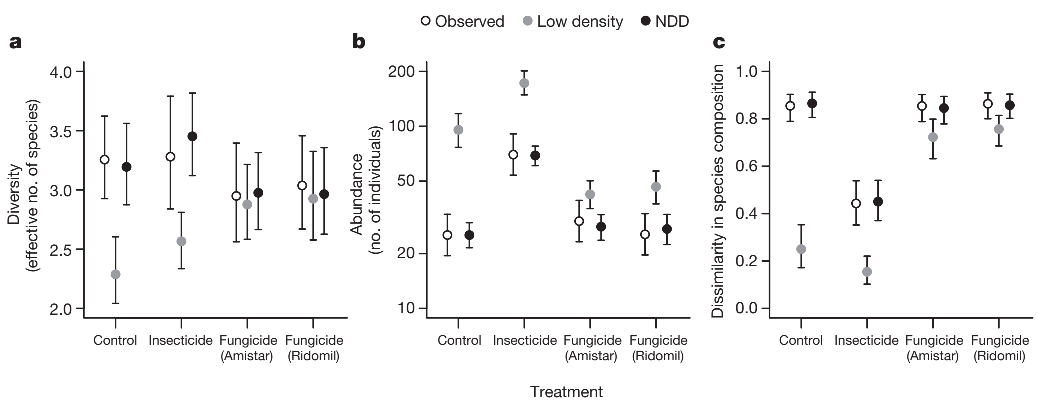


Figure 4 | Including NDD in model simulations reproduces the observed diversity patterns, whereas excluding NDD underestimates diversity in the control and insecticide treatments. Observed diversity (**a**), total abundance (**b**) and dissimilarity in species composition to the seeds (**c**) in each treatment were compared with those simulated either assuming a constant survival for each species equal to its estimated low-density survival (low-density survival) or NDD survival. The error bars are 95% confidence intervals of the mean extracted from models fitted to the data (observed) or the 95% quantile (simulations) from 1,000 simulations in each scenario.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 July; accepted 26 November 2013.

Published online 22 January 2014.

1. Gibson, L. *et al.* Primary forests are irreplaceable for sustaining tropical biodiversity. *Nature* **478,** 378–381 (2011).
2. Wright, S. J. Plant diversity in tropical forests: a review of mechanisms of species coexistence. *Oecologia* **130,** 1–14 (2001).
3. Janzen, D. H. Herbivores and the number of tree species in tropical forests. *Am. Nat.* **104,** 501–528 (1970).
4. Connell, J. H. in *Dynamics of Numbers in Populations* (eds den Boer, P. J. & Gradwell, G. R.) 298–312 (PUDOC, 1971).
5. Harms, K. E., Wright, S. J., Calderón, O., Hernández, A. & Herre, E. A. Pervasive density-dependent recruitment enhances seedling diversity in a tropical forest. *Nature* **404,** 493–495 (2000).
6. Metz, M. R., Sousa, W. & Valencia, R. Widespread density-dependent seedling mortality promotes species coexistence in a highly diverse Amazonian rain forest. *Ecology* **91,** 3675–3685 (2010).
7. Bagchi, R. *et al.* Testing the Janzen-Connell mechanism: pathogens cause overcompensating density dependence in a tropical tree. *Ecol. Lett.* **13,** 1262–1269 (2010).
8. Comita, L. S. & Hubbell, S. P. Local neighborhood and species' shade tolerance influence survival in a diverse seedling bank. *Ecology* **90,** 328–334 (2009).
9. Terborgh, J. Enemies maintain hyperdiverse tropical forests. *Am. Nat.* **179,** 303–314 (2012).
10. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31,** 343–366 (2000).
11. Curran, L. M. *et al.* Lowland forest loss in protected areas of Indonesian Borneo. *Science* **303,** 1000–1003 (2004).
12. Achard, F. *et al.* Determination of deforestation rates of the world's humid tropical forests. *Science* **297,** 999–1002 (2002).
13. Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* **320,** 1444–1449 (2008).
14. Bell, T., Freckleton, R. P. & Lewis, O. T. Plant pathogens drive density-dependent seedling mortality in a tropical tree. *Ecol. Lett.* **9,** 569–574 (2006).
15. Mangan, S. A. *et al.* Negative plant-soil feedback predicts tree-species relative abundance in a tropical forest. *Nature* **466,** 752–755 (2010).
16. Bever, J. D. Feedback between plants and their soil communities in an old field community. *Ecology* **75,** 1965–1977 (1994).
17. Packer, A. & Clay, K. Soil pathogens and spatial patterns of seedling mortality in a temperate tree. *Nature* **404,** 278–281 (2000).
18. Webb, C. O. & Peart, D. R. Seedling density dependence promotes coexistence of Bornean rain forest trees. *Ecology* **80,** 2006–2017 (1999).
19. Theimer, T. C., Gehring, C. A., Green, P. T. & Connell, J. H. Terrestrial vertebrates alter seedling composition and richness but not diversity in an Australian tropical rain forest. *Ecology* **92,** 1637–1647 (2011).
20. Leigh, E. G., Wright, S. J., Herre, E. A. & Putz, F. E. The decline of tree diversity on newly isolated tropical islands: a test of a null hypothesis and some implications. *Evol. Ecol.* **7,** 76–102 (1993).
21. Terborgh, J. *et al.* Tree recruitment in an empty forest. *Ecology* **89,** 1757–1768 (2008).
22. Hammond, D. S. & Brown, V. K. in *Dynamics of Tropical Communities* (eds Newbery, D. M., Prins, H. H. T. & Brown, N. D.) 51–78 (Blackwell, 1998).
23. Horn, H. S. Measurement of "overlap" in comparative ecological studies. *Am. Nat.* **100,** 419–424 (1966).
24. Comita, L. S., Muller-Landau, H. C., Aguilar, S. & Hubbell, S. P. Asymmetric density dependence shapes species abundances in a tropical tree community. *Science* **329,** 330–332 (2010).
25. Kobe, R. K. & Vriesendorp, C. F. Conspecific density dependence in seedlings varies with species shade tolerance in a wet tropical forest. *Ecol. Lett.* **14,** 503–510 (2011).
26. Bagchi, R. *et al.* Impacts of logging on density-dependent predation of dipterocarp seeds in a southeast Asian rainforest. *Phil. Trans. R. Soc. B* **366,** 3246–3255 (2011).
27. Paine, C. E. T. & Beck, H. Seed predation by neotropical rain forest mammals increases diversity in seedling recruitment. *Ecology* **88,** 3076–3087 (2007).
28. Norghauer, J. M., Malcolm, J., Zimmerman, B. & Felfili, J. An experimental test of density- and distant-dependent recruitment of mahogany (*Swietenia macrophylla*) in southeastern Amazonia. *Oecologia* **148,** 437–446 (2006).
29. Bridgewater, S. G. M. *et al.* A preliminary checklist of the vascular plants of the Chiquibul Forest, Belize. *Edinb. J. Bot.* **63,** 269–321 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements Permission to undertake research in the Chiquibul Forest Reserve was granted by the Ministry of Natural Resources, Belize under Scientific Collection/Research Permit CD/60/3/07(20). We thank the staff at the Las Cuevas Research Station (the late N. Bol, C. Bol, M. Bol and J. Boucher) for their help; and R. Cocomb, E. Miles, C. Rasell, M. Senior, T. Swinfield and O. Theisinger provided field assistance. H. Rue provided advice on implementing measurement error models in INLA. This research was funded by the Natural Environment Research Council (NERC; standard grant NE/DO10721/1) and S.G. was funded by grant 126296 from the Academy of Finland.

Author Contributions O.T.L., R.P.F. and S.J.G. conceived the project and obtained funding. R.B., R.E.G., S.G., O.T.L., L.N. and C.E.A. established fieldwork design and protocols, and carried out the fieldwork with advice from R.P.F. and S.J.G. Data analysis was carried out by R.B. with input from R.P.F. and O.T.L. R.B. wrote the first draft of the manuscript and all authors contributed to discussing the results and editing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to O.T.L. (owen.lewis@zoo.ox.ac.uk).

METHODS

Field survey. Our field site was close to the Las Cuevas Research Station in southwest Belize (16° 43′ 53′′ N, 88° 59′ 11′′ W) at 450 m elevation within the 170,000-ha Chiquibul Forest Reserve protected area. This site has limestone geology and a relatively intact flora and vertebrate fauna. It experiences a marked dry season, typically from February to May, with annual rainfall approximately 1,500 to 1,800 mm (ref. 30). We established 36 sampling stations on the forest floor, positioned at 20-m intervals on a 120 m × 120 m grid. Each station comprised seven 1-m^2 quadrats, placed as close together as possible while avoiding trees and large rocks. Three of the quadrats at each station were randomly selected as locations for 1-m^2 seed traps made from 1-mm mesh fibreglass netting, suspended 80 cm above the ground using PVC poles. The remaining quadrats were cleared of existing seedlings and assigned at random to one of three enemy exclusion treatments or to a control treatment. Two fungicide treatments were used: Amistar (Syngenta Ltd; active ingredient, azoxystrobin), which has broad-spectrum systemic activity against a range of plant pathogenic fungi, and Ridomil Gold MZ 68WP (Syngenta Ltd; active ingredients, mancozeb and metalxyl), which protects plants from infection by oomycetes and fungi. The insecticide used was Engeo (Syngenta Ltd; active ingredient, thiamethoxam), which provides both systemic and contact protection against a range of insects. Pesticides were applied weekly with a hand mister, following the manufacturers' guidelines (0.005 g of Amistar, 0.25 g of Ridomil Gold or 0.0025 ml of Engeo, in each case dissolved in 50 ml of water). Control plots were sprayed with 50 ml of water at the same time as pesticide applications. Treatments began in July 2007 except for the Engeo application, which began in April 2008. All treatments were applied weekly until September 2009. Only data from April 2008 onwards (during which all treatments were applied) were used in the analyses presented here.

Seeds were collected weekly from the traps; damaged or inviable seeds (partially eaten or immature) were discarded and the remaining seeds were counted and identified to species level, where possible, or as morphospecies. A subset of the seeds from each species and morphospecies were placed on moist tissue paper in seed germination trays. We photographed examples of all seed and seedling morphospecies to match seeds to seedlings in cases where species identification was not possible. This ensured consistent classification throughout the experiment, and facilitated subsequent plant identification. In this way we matched 97% of the individual seeds collected in our study to seedlings.

We censused the seedling plots for new seedlings every month during the peak period of fruiting and recruitment (April to August) and less frequently (every 2 to 4 months) during the rest of the year. At each census, all new seedlings were tagged and identified with species or morphospecies. Unidentified seedlings were photographed. By comparing these photographs to seedlings germinated from collected seeds we were able to match 90% of the observed seedlings to seeds.

To confirm that the significant effects of insecticide treatment were a consequence of reduced attack from insects rather than a direct effect of Engeo on plant survival[31], we set up experiments in May 2010 in which a subset of the focal plant species (*Stemmadenia donnell-smithii* ($n = 60$), *Cordia alliodora* ($n = 80$), *Cryosophila stauracantha* ($n = 70$), *Combretum laxum/fruticosum* ($n = 70$), *Terminalia amazonia* ($n = 100$) and *Forsteronia* sp. ($n = 70$)) were grown from seed at high density under insect-free conditions, and with Engeo treatment. Freshly collected seeds were sown into 60 seed trays (36 cm length × 24 cm width × 5 cm depth) filled with locally collected soil that had been sorted to remove large stones and roots. Each tray was divided into six sections, with seeds of each species sown into one section. Trays were enclosed in a bag made from insect-proof nylon netting to exclude insects. The netting was raised above the surface of the tray to allow seedlings to grow. For the shadehouse experiment, 30 trays were placed in randomly allocated positions on raised benches in a small forest gap, covered with waterproof shade netting. For the field experiment 30 trays were placed on the forest floor in a randomized grid design, spaced by 200 cm. Trays in the shadehouse were watered regularly (approximately every 2 to 3 days). Half of the trays (chosen at random) in both experiments were sprayed weekly with 0.0025 ml m^{-2} of Engeo using a hand mister. The remaining trays were sprayed with an equal volume of water. Germinating and surviving seedlings were censused after 8 weeks (shadehouse experiment) or 7 weeks (field experiment). We analysed the number of seedlings at the end of the experiment as a function of insecticide treatment using generalized linear models for each species, assuming a negative binomial distribution for the errors. No significant ($P < 0.05$) effects of Engeo on survivorship were documented in any species in either experiment (see Extended Data Table 2).

Analysis. We calculated the total number of seeds or seedlings observed in each seedling plot (N) and the reciprocal of the Simpson's dominance index ($1/D$, $D = \sum_k^s p_k^2$, where p_k is the proportional abundance of species k in a community with s species) as a measure of the effective number of species[32]. We quantified differences in species composition among treatments by calculating the Morisita–Horn index of dissimilarity (R_h)[23]

$$R_{hij} = 1 - \frac{2 \cdot \sum_k^s \left(N_{ik} \cdot N_{jk}\right)}{\left(\frac{\sum_k^s N_{ik}^2}{\left(\sum_k^s N_{ik}\right)^2} + \frac{\sum_k^s N_{jk}^2}{\left(\sum_k^s N_{jk}\right)^2}\right) \cdot \sum_k^s N_{ik} \cdot \sum_k^s N_{jk}} \tag{1}$$

between the seed traps and all the seedling plots, i, at each station, j. We calculated the pairwise dissimilarity between each plot and each trap, and then took the mean for the three traps at each station. Results were qualitatively unchanged using other diversity and dissimilarity metrics (for example, Shannon's diversity index and the Bray–Curtis index of dissimilarity; see Extended Data Tables 3 and 4). All diversity and dissimilarity indices were calculated using the 'vegan' package[33] in R v3.0.1 (ref. 34). We compared these metrics between control plots and seed traps and among pesticide treatments (control, insect exclusion with Engeo, true fungi exclusion with Amistar or oomycete and true fungi exclusion with Ridomil), using linear mixed-effects models (fitted using the 'nlme' package[35] in R v3.0.1 (ref. 34)) with different intercepts for the stations included as a normally distributed random effect. We assumed a Gaussian error distribution for the models of N (log-transformed), $1/D$ (log-transformed) and dissimilarity (logit-transformed). There was evidence of heteroscedasticity in the residuals of the models of $1/D$ and R_h, so this was accounted for by explicitly modelling the variance as a function of pesticide treatment (for $1/D$) or as an exponential function of the expected values (for R_h).

We used these models to test three hypotheses: first, diversity is greater among seedlings than among seeds; second, excluding natural enemies with pesticides decreases diversity; and, third, excluding natural enemies with pesticides alters species composition.

For a subset of species we examined the effects of pesticides, seed density and their interaction on seedling recruitment at the level of individual species. For this analysis, we selected all 18 species that met two criteria: seeds and/or seedlings of these species were recorded at ≥5 stations (the sets of stations with seeds and seedlings did not have to overlap); and mean seed density varied at least threefold among stations. The species that met these criteria are listed in Supplementary Table 1. The relationship between the number of seeds in plot i at station j, $N_{0,ij}$, and the expected number of recruits, $N_{1,ij}$ can be described by the equation[5]

$$N_{1,ij} = \exp(\alpha) N_{0,ij}^{\beta} \tag{2}$$

where $\exp(\alpha)$ is the ratio of seedlings to seeds at low density ($N_{0,ij} = 1$). The parameter β is 1 if survival is independent of conspecific density and less than 1 when this ratio is reduced at high density (that is, NDD). Because we did not measure the seed rain in the seedling plots at each station j directly, $N_{0,ij}$ has to be estimated from the adjacent seed traps at station j instead. Ignoring the error in these estimates of $N_{1,ij}$ biases the estimation of β towards 0 (ref. 36), and therefore overestimates the importance of NDD. To overcome this potential bias, we modelled N_0 and N_1 jointly as:

$$\hat{N}_{0,ij} = \mathrm{NegBin}\left(\lambda_j, \kappa_0\right); \lambda_j \sim \mathrm{lognorm}\left(\bar{\lambda}, \sigma^2\right)$$
$$\hat{N}_{1,ijL} = \mathrm{NegBin}(\exp(\alpha_L) N_{0,ij}^{\beta_L}, \kappa_{1,L}) \tag{3}$$

where both $\hat{N}_{0,ij}$ (the number of seeds in plot i at station j) and $\hat{N}_{1,ij}$ (the number of recruits in plot i at station j) were drawn from negative binomial distributions defined by the expected number of individuals and stage ($t = 0, 1$) and treatment (L) specific size or overdispersion parameters, $\kappa_{t,L}$. The expected number of seeds in the plots at station j is λ_j and the λ_j were drawn from a lognormal distribution with mean $\bar{\lambda}$ and variance σ^2. The number of seeds falling in the seedling plots is treated as missing data which need to be imputed from the seed trap data collected at the same station. The parameters α_L and β_L correspond to the low-density survival rate and effect of density on recruitment under treatment L as described by equation (2). This hierarchical model was fitted using the INLA package[37] in R v3.0.1 (ref. 34).

We used estimates of β_L from these models to test two hypotheses for each species: first, survival is negatively density dependent ($\beta_{\mathrm{control}} < 1$); second, natural enemies cause the observed negative density dependence, so that applying pesticides weakens the relationship between seed density and survival ($\beta_{\mathrm{control}} < \beta_{\mathrm{pesticide}}$).

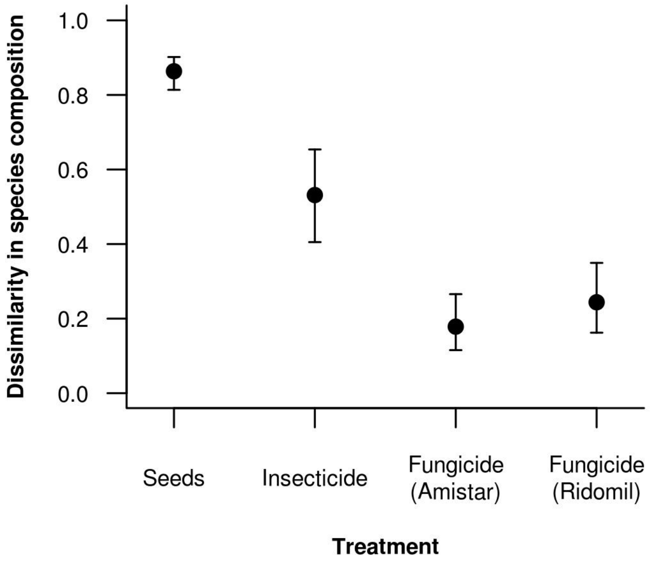
We tested whether the estimates of β across the 18 species were significantly different from 1 in the control treatment and whether they varied among pesticide treatments and with the logarithm of the seed abundance ($N_{0,k}$) of each species, k. This was achieved by fitting a linear mixed-effects model to estimates of β with species included as a random effect. The contribution of each estimate of β to the model was weighted by $\omega_{k,L}$, the inverse of its standard deviation. The model can be described as

$$\hat{\beta}_{kL} = \gamma_0 + \gamma_{1,L} + \gamma_2 \cdot \log(N_{0.k}) + \gamma_{3,L} \cdot \log(N_{0,k}) + b_k + \omega_{k,L}\varepsilon_{kL}$$
$$b_k \sim \mathrm{Norm}\left(0,\sigma_b^2\right);\ \varepsilon_{kL} \sim \mathrm{Norm}(0,\sigma_\varepsilon^2) \quad (4)$$

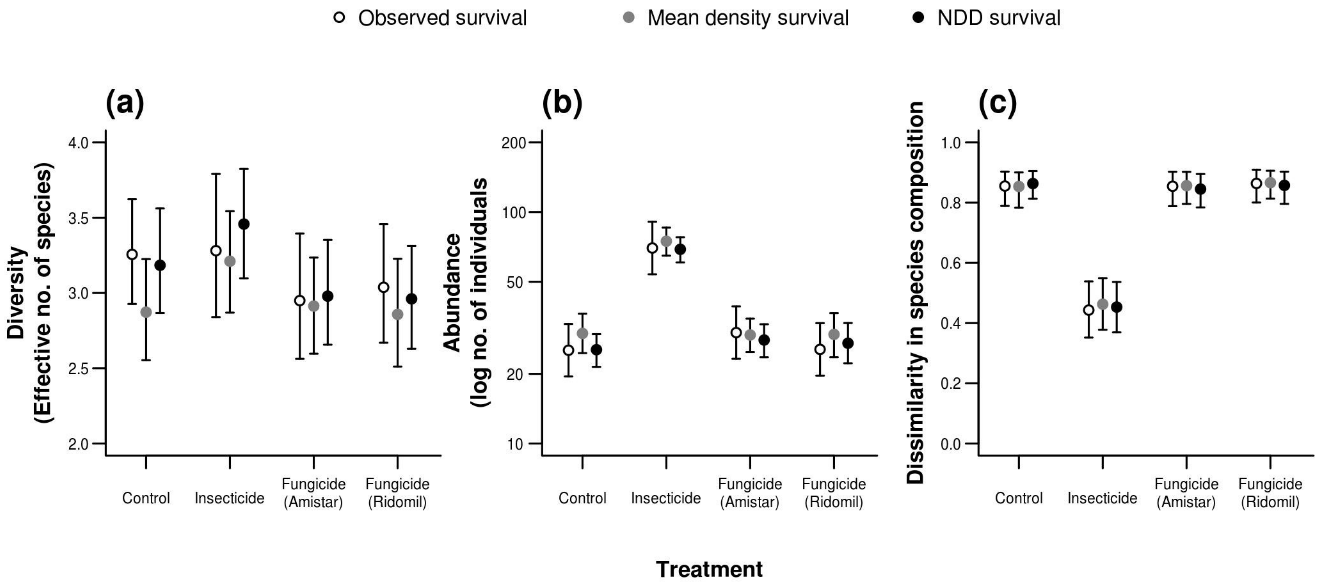
where the γ represent the estimated fixed effects parameters, b_k is the random effect for species k and ε_{kL} is the error under treatment L for species k. The parameters $\gamma_{1,k}$ (pesticide effect on mean NDD) and $\gamma_{3,k}$ (pesticide effect on the relationship between overall seed abundance and strength of NDD) are zero for the control treatment and represent the change in these parameters under each pesticide treatment compared to the control.

Finally, we tested whether the estimated effects of NDD and pesticides on recruitment of individual species could account for the observed differences in diversity among pesticide treatments. We used the parameters estimated in the 18 species-specific models to simulate new communities in three scenarios. For each species, the number of seeds in each trap was drawn from a negative binomial distribution with mean λ_j and size κ_0. In the 'NDD survival' scenario, the number of seedlings in plots at station j with treatment L was drawn from a negative binomial distribution with mean $= \exp(\alpha_{1,L})\lambda_j^{\beta_{1,L}}$ and size $= \kappa_{1,L}$. The 'low-density survival' scenario assumed that survival was independent of seed density and was equal to the survival for each species within each pesticide treatment when density was 1 (that is, when seedlings had no conspecific neighbours). The number of seedlings was therefore drawn from a negative binomial distribution with mean $= \exp(\alpha_{1,L})\lambda_j$. We then calculated the effective number of species for the simulated communities at each station and treatment combination and extracted the means for each treatment. This procedure was repeated 1,000 times and the median and 95% quantiles across the simulations were extracted in each scenario. A similar procedure was used to simulate communities expected in a third scenario, consistent with previous studies[5], where seed-to-seedling transition probabilities reflected those recorded at the mean seed density for each species. This was achieved by refitting the model to each species after fixing the values of all the β_L to 1 and using this model for the simulations. We then calculated the mean total abundance, effective number of species and dissimilarity to species composition of the seeds in each treatment using the observed data for the 18 species. We compared these observed data to the simulations in the low-density and NDD scenarios (main text) and the mean-density and NDD scenarios (Extended Data Fig. 2).

30. Bridgewater, S. *A Natural History of Belize* (Univ. Texas Press, 2012).
31. Ford, K. A. *et al.* Neonicotinoid insecticides induce salicylate-associated plant defense responses. *Proc. Natl Acad. Sci. USA* **107,** 17527–17532 (2010).
32. Jost, L. Entropy and diversity. *Oikos* **113,** 363–375 (2006).
33. Oksanen, J. *et al.* vegan: community ecology package v.2.0-8 (R Foundation for Statistical Computing, 2013).
34. R Development Core Team. *R: a language and environment for statistical computing v.3.0.1* (R Foundation for Statistical Computing, 2013).
35. Pinheiro, J. C. & Bates, D. M. *Mixed-Effects Models in S and S-Plus* (Springer, 2000).
36. Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. *Measurement Error in Nonlinear Models: A Modern Perspective* 2nd edn (Chapman & Hall/CRC, 2006).
37. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B* **71,** 319–392 (2009).



Extended Data Figure 1 | The mean abundance-weighted Morisita–Horn dissimilarity in species composition (and 95% confidence intervals), comparing seedlings recruiting in the control plots with seedlings in the pesticide treatments and with seeds falling into seed traps.



Extended Data Figure 2 | A comparison of the observed seedling communities (observed survival) with those simulated either fixing survival to the mean for each species in each treatment (mean density survival) or allowing survival to be negatively density dependent (NDD survival). The simulated values are means and 95% confidence intervals based on 1,000 simulations for effective number of species, total abundance and community dissimilarity to seeds falling in adjacent traps.

Extended Data Table 1 | Coefficients from the model relating the strength of NDD to treatment, log total seed abundance, and their interaction

Term	Parameter	Std. Error	d. f.	t	P
Intercept (Water) (γ_0)	-0.12	0.044	48	-2.68	0.010
Insecticide effect (Engeo) ($\gamma_{1,\ Engeo}$)	0.00	0.047	48	-0.02	0.982
Fungicide effect (Amistar) ($\gamma_{1,\ Amistar}$)	0.05	0.048	48	1.02	0.312
Fungicide effect (Ridomil) ($\gamma_{1,\ Ridomil}$)	-0.02	0.049	48	-0.34	0.732
log N_0 (standardised) (γ_2)	-0.37	0.086	16	-4.33	0.001
Engeo:log N_0 (standardised) ($\gamma_{3,\ Engeo}$)	0.23	0.085	48	2.65	0.010
Amistar:log N_0 (standardised) ($\gamma_{3,\ Amistar}$)	0.25	0.096	48	2.60	0.012
Ridomil:log N_0 (standardised) ($\gamma_{3,\ Ridomil}$)	0.28	0.095	48	3.00	0.004
Among species variance (σ_b^2)	0.012				
Residual variance (σ_ε^2)	0.469				

The strength of NDD was measured as the coefficient for seed density, β, from equation (3). The log total seed abundance was standardized across species. The model was fitted as a mixed-effects model with the contribution of each value of β weighted by the inverse of its standard deviation.

Extended Data Table 2 | Coefficients from the negative binomial model fitted to the shadehouse and field trials of effects of the insecticide Engeo on seedling survival

		Shadehouse				Field			
		Coefficient	SE	Z	P	Coefficient	SE	Z	P
Stemmadenia	Intercept	-0.57	0.044	-12.858	<0.001	-0.67	0.128	-5.220	<0.001
	EngeoTreatment	0.04	0.062	0.621	0.535	-0.05	0.181	-0.282	0.778
Cordia	Intercept	-0.43	0.086	-5.053	<0.001	-0.35	0.064	-5.495	<0.001
	EngeoTreatment	-0.07	0.121	-0.592	0.554	-0.06	0.090	-0.633	0.527
Cryosophila	Intercept	-0.33	0.036	-8.941	<0.001	-0.61	0.042	-14.366	<0.001
	EngeoTreatment	-0.01	0.051	-0.154	0.877	-0.03	0.060	-0.531	0.595
Combretum	Intercept	-0.67	0.043	-15.452	<0.001	-3.11	0.332	-9.364	<0.001
	EngeoTreatment	0.02	0.061	0.243	0.808	0.83	0.456	1.824	0.068
Terminalia	Intercept	-4.83	0.289	-16.725	<0.001	-6.62	0.743	-8.908	<0.001
	EngeoTreatment	-0.09	0.417	-0.208	0.835	1.10	0.878	1.251	0.211
Forsteronia	Intercept	-0.67	0.043	-15.567	<0.001	-1.12	0.108	-10.360	<0.001
	EngeoTreatment	0.01	0.061	0.152	0.879	0.00	0.152	0.000	1.000

Note that Engeo did not have a significant effect on survival in any of the species tested (shaded rows).

Extended Data Table 3 | Tests of pesticide effects on seedling species diversity using different diversity indices

Treatment	Parameter	Std.Error	d. f.	t	P
Log Inverse Simpson's Diversity Index (1/D)					
Intercept (Water)	1.30	0.059	105	22.04	0.000
Insecticide (Engeo)	-0.06	0.077	105	-0.83	0.410
Fungicide (Amistar)	-0.18	0.072	105	-2.45	0.016
Fungicide (Ridomil)	-0.08	0.072	105	-1.10	0.274
Log Shannon's Diversity Index (H)					
Intercept (Water)	1.56	0.053	105	29.30	0.000
Insecticide (Engeo)	0.03	0.063	105	0.54	0.591
Fungicide (Amistar)	-0.15	0.067	105	-2.29	0.024
Fungicide (Ridomil)	-0.07	0.067	105	-1.05	0.298
Log Fisher's alpha (α)					
Intercept (Waterl)	1.33	0.084	104	15.92	0.000
Insecticide (Engeo)	-0.13	0.086	104	-1.57	0.120
Fungicide (Amistar)	-0.18	0.095	104	-1.88	0.063
Fungicide (Ridomil)	-0.04	0.126	104	-0.33	0.744
Rarefied Species Richness					
Intercept (Waterl)	1.16	0.033	105	35.38	0.000
Insecticide (Engeo)	-0.06	0.037	105	-1.56	0.122
Fungicide (Amistar)	-0.11	0.043	105	-2.55	0.012
Fungicide (Ridomil)	-0.03	0.041	105	-0.77	0.446

The indices were calculated using the vegan v2.0833 package in R v3.0.134. Rarefaction was based on samples of five individuals. The table presents parameters of linear mixed-effects models fitted to these metrics as functions of treatment. The first parameter (intercept) represents mean diversity in the control treatment and the following parameters represent the difference between the control treatment and the three pesticide treatments.

Extended Data Table 4 | Tests of pesticide effects on dissimilarity in species composition, comparing assemblages of seedlings germinating in plots to those of seeds falling in adjacent seed traps

Treatment	Parameter	Std.Error	d. f.	t	P
Logit Morisita-Horn Dissimilarity Index					
Intercept (Water)	1.72	0.240	105	7.18	0.000
Insecticide (Engeo)	-1.91	0.243	105	-7.86	0.000
Fungicide (Amistar)	-0.03	0.270	105	-0.12	0.904
Fungicide (Ridomil)	0.12	0.272	105	0.44	0.664
Binomial Dissimilarity Index					
Intercept (Water)	8.251	0.205	105	40.208	0.000
Insecticide (Engeo)	-0.965	0.183	105	-5.278	0.000
Fungicide (Amistar)	0.070	0.205	105	0.342	0.733
Fungicide (Ridomil)	0.013	0.203	105	0.063	0.950
Logit Bray-Curtis Dissimilarity Index					
Intercept (Water)	3.123	0.175	105	17.896	0.000
Insecticide (Engeo)	-1.412	0.110	105	-12.889	0.000
Fungicide (Amistar)	-0.144	0.110	105	-1.312	0.192
Fungicide (Ridomil)	-0.008	0.110	105	-0.075	0.940
Logit Jaccard Dissimilarity Index					
Intercept (Water)	3.814	0.175	105	21.757	0.000
Insecticide (Engeo)	-1.417	0.110	105	-12.847	0.000
Fungicide (Amistar)	-0.148	0.110	105	-1.344	0.182
Fungicide (Ridomil)	-0.008	0.110	105	-0.075	0.940

Four alternative metrics of dissimilarity were calculated using the vegan 2.08 package in R v3.0.1 (ref. 33). The table presents the fixed-effects parameters of linear mixed-effects models used to describe these metrics as functions of treatment. The intercept represents the mean dissimilarity between seed and control plot assemblages, and the other three parameters indicate the difference between seed–control plot dissimilarity and seed–pesticide plot dissimilarity.

LETTER

doi:10.1038/nature12872

Three keys to the radiation of angiosperms into freezing environments

Amy E. Zanne[1,2], David C. Tank[3,4], William K. Cornwell[5,6], Jonathan M. Eastman[3,4], Stephen A. Smith[7], Richard G. FitzJohn[8,9], Daniel J. McGlinn[10], Brian C. O'Meara[11], Angela T. Moles[6], Peter B. Reich[12,13], Dana L. Royer[14], Douglas E. Soltis[15,16,17], Peter F. Stevens[18], Mark Westoby[9], Ian J. Wright[9], Lonnie Aarssen[19], Robert I. Bertin[20], Andre Calaminus[15], Rafaël Govaerts[21], Frank Hemmings[6], Michelle R. Leishman[9], Jacek Oleksyn[12,22], Pamela S. Soltis[16,17], Nathan G. Swenson[23], Laura Warman[6,24] & Jeremy M. Beaulieu[25]

Early flowering plants are thought to have been woody species restricted to warm habitats[1–3]. This lineage has since radiated into almost every climate, with manifold growth forms[4]. As angiosperms spread and climate changed, they evolved mechanisms to cope with episodic freezing. To explore the evolution of traits underpinning the ability to persist in freezing conditions, we assembled a large species-level database of growth habit (woody or herbaceous; 49,064 species), as well as leaf phenology (evergreen or deciduous), diameter of hydraulic conduits (that is, xylem vessels and tracheids) and climate occupancies (exposure to freezing). To model the evolution of species' traits and climate occupancies, we combined these data with an unparalleled dated molecular phylogeny (32,223 species) for land plants. Here we show that woody clades successfully moved into freezing-prone environments by either possessing transport networks of small safe conduits[5] and/or shutting down hydraulic function by dropping leaves during freezing. Herbaceous species largely avoided freezing periods by senescing cheaply constructed aboveground tissue. Growth habit has long been considered labile[6], but we find that growth habit was less labile than climate occupancy. Additionally, freezing environments were largely filled by lineages that had already become herbs or, when remaining woody, already had small conduits (that is, the trait evolved before the climate occupancy). By contrast, most deciduous woody lineages had an evolutionary shift to seasonally shedding their leaves only after exposure to freezing (that is, the climate occupancy evolved before the trait). For angiosperms to inhabit novel cold environments they had to gain new structural and functional trait solutions; our results suggest that many of these solutions were probably acquired before their foray into the cold.

Flowering plants (angiosperms) today grow in a vast range of environmental conditions, with this breadth probably related to their diverse morphology and physiology[7]. However, early angiosperms are generally thought to have been woody and restricted to warm understory habitats[1–3]. Debate continues about these assertions, in part because of the paucity of fossils and uncertainty in reconstructing habits for these first representatives[8–11]. Nevertheless, greater mechanical strength of woody tissue would have made extended lifespans possible at a height necessary to compete for light[12,13]. A major challenge resulting from increased stature is that hydraulic systems must deliver water at tension to greater heights: as path lengths increase so too does resistance[5]. Among extant strategies, the most efficient method of water delivery is through large-diameter water-conducting conduits (that is, vessels and tracheids) within xylem[5].

Early in angiosperm evolution they probably evolved larger conduits for water transport, especially compared with their gymnosperm cousins[14]. Although efficient in delivering water, these larger cells would have impeded angiosperm colonization of regions characterized by episodic freezing[14,15], as the propensity for freezing-induced embolisms (air bubbles produced during freeze/thaw events that block hydraulic pathways)

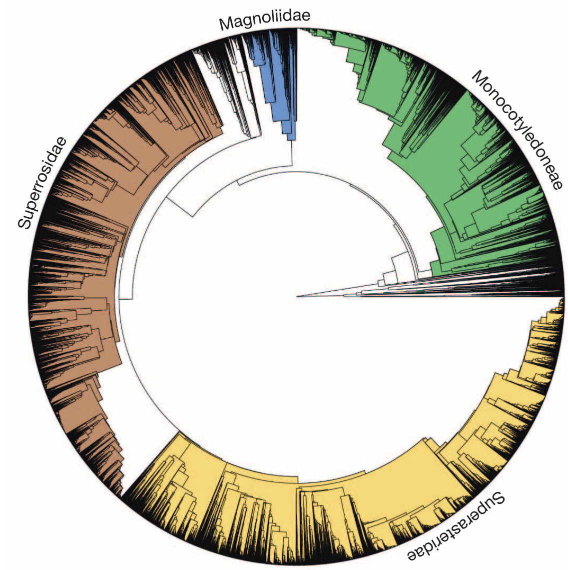


Figure 1 | Time-calibrated maximum-likelihood estimate of the molecular phylogeny for 31,749 species of seed plants. The four major angiosperm lineages discussed in the text are highlighted: Monocotyledoneae (green), Magnoliidae (blue), Superrosidae (brown) and Superasteridae (yellow). Non-seed plant outgroups (that is, bryophytes, lycophytes and monilophytes) were removed for the purposes of visualization.

[1]Department of Biological Sciences, George Washington University, Washington DC 20052, USA. [2]Center for Conservation and Sustainable Development, Missouri Botanical Garden, St Louis, Missouri 63121, USA. [3]Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844, USA. [4]Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho 83844, USA. [5]Department of Ecological Sciences, Systems Ecology, de Boelelaan 1085, 1081 HV Amsterdam, the Netherlands. [6]Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia. [7]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA. [8]Department of Zoology and Biodiversity Research Centre, University of British Columbia, Vancouver, British Columbia V6T1Z4, Canada. [9]Department of Biological Sciences, Macquarie University, Sydney, New South Wales 2109, Australia. [10]Department of Biology and the Ecology Center, Utah State University, Logan, Utah 84322, USA. [11]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee 37996, USA. [12]Department of Forest Resources, University of Minnesota, St Paul, Minnesota 55108, USA. [13]Hawkesbury Institute for the Environment, University of Western Sydney, Penrith, New South Wales 2751, Australia. [14]Department of Earth and Environmental Sciences, Wesleyan University, Middletown, Connecticut 06459, USA. [15]Department of Biology, University of Florida, Gainesville, Florida 32611, USA. [16]Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, USA. [17]Genetics Institute, University of Florida, Gainesville, Florida 32611, USA. [18]Department of Biology, University of Missouri—St Louis, St Louis, Missouri 63121, USA. [19]Department of Biology, Queen's University, Kingston, Ontario K7L 3N6, Canada. [20]Department of Biology, College of the Holy Cross, Worcester, Massachusetts 01610, USA. [21]Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, United Kingdom. [22]Polish Academy of Sciences, Institute of Dendrology, 62-035 Kornik, Poland. [23]Department of Plant Biology and Ecology, Evolutionary Biology and Behavior, Program, Michigan State University, East Lansing, Michigan 48824, USA. [24]Institute of Pacific Islands Forestry, USDA Forest Service, Hilo, Hawaii 96720, USA. [25]National Institute for Mathematical & Biological Synthesis, University of Tennessee, Knoxville, Tennessee 37996, USA.

increases as conduit diameter increases[5]. Three evolutionary solutions seemingly arose to address the challenges of freezing: (1) woody species withstood freezing temperatures without serious loss of hydraulic function by building safe water-transport networks consisting of small-diameter conduits; (2) woody species shut down hydraulic function by becoming deciduous, dropping leaves during freezing periods; and (3) herbaceous species largely avoided freezing by senescing cheaply constructed aboveground tissue and overwintering, probably as seeds or underground storage organs. However, the order in which angiosperms are likely to have acquired these solutions relative to exposure to and persistence in the cold[16] remains unclear.

Proportions of herbaceous species, deciduous species and those with small water-conducting conduits increase towards the poles[1,4,17,18], and an earlier limited survey of angiosperm families indicated that herbaceousness and ability to cope with freezing evolved in parallel[19]. However, exactly how global-scale ecological patterns are linked to functional evolution of angiosperms is uncertain. We dissect the contributions of different evolutionary solutions allowing angiosperms to cope with periodic freezing and assess likely pathways by which clades acquired these traits (that is, timing of evolution in climate occupancy relative to trait evolution).

We compiled a very large species-level database of angiosperm growth habits (49,064 species, which is 16.4% of accepted land plant species[20] in The Plant List; http://www.theplantlist.org), leaf phenology, conduit diameter and freezing climate exposure. To trace species trait and climate occupancy relationships over evolutionary time, we generated an unparalleled time-scaled molecular phylogeny for 32,223 land plant species in our database (Fig. 1; http://www.onezoom.org/vascularplants_tank2013nature.htm). This timetree gives us the most comprehensive view yet into the evolutionary history of angiosperms. On the basis of their geographic distributions, we classified species' climate occupancies with respect to freezing: 'freezing unexposed', only encountering temperatures >0 °C across a species' range; and 'freezing exposed', encountering temperatures ≤0 °C somewhere across a species' range. This dichotomy assumes that climate tracking through environmental changes is more common than the evolution of climate occupancy; this is more likely to be true if freezing exposure has a physiological cost in regions without freezing[21]. Species were further distinguished by leaf phenology (deciduous or evergreen); conduit diameter (large ≥0.044 mm, or small <0.044 mm; as 0.044 mm diameter is the diameter above which freezing-induced embolisms are believed to become frequent at modest tensions[22]); and growth form (woody or herbaceous, with woody species defined as those maintaining a prominent aboveground stem that is persistent over time and with changing environmental conditions; see Extended Data Fig. 1 for examples of angiosperms with woody growth habits as we define them, and Extended Data Table 1 for a breakdown of growth habit by order within angiosperms).

Among woody species we asked whether evolutionary transitions between climate occupancy states were significantly associated with shifts in leaf phenology and/or conduit diameter. Among all angiosperms we asked whether evolutionary transitions between climate occupancy states were significantly associated with shifts in growth form. We determined the relative lability of climate occupancy (exposure to freezing) versus traits (growth form, leaf phenology or conduit diameter) by summing all climate occupancy transitions and dividing by the sum of all trait transitions. We also devised a novel summary based on these evolutionary transition rates that provides the likeliest pathways from the purported early angiosperm (woody, evergreen, with large conduits and freezing unexposed) to a plant with traits for freezing conditions. Because evolutionary rates are unlikely to be uniform at this phylogenetic scale, we ran growth form analyses both across the entire angiosperm data set and also within each of four major lineages: Monocotyledoneae (monocots), Magnoliidae (magnoliids), Superrosidae (superrosids)

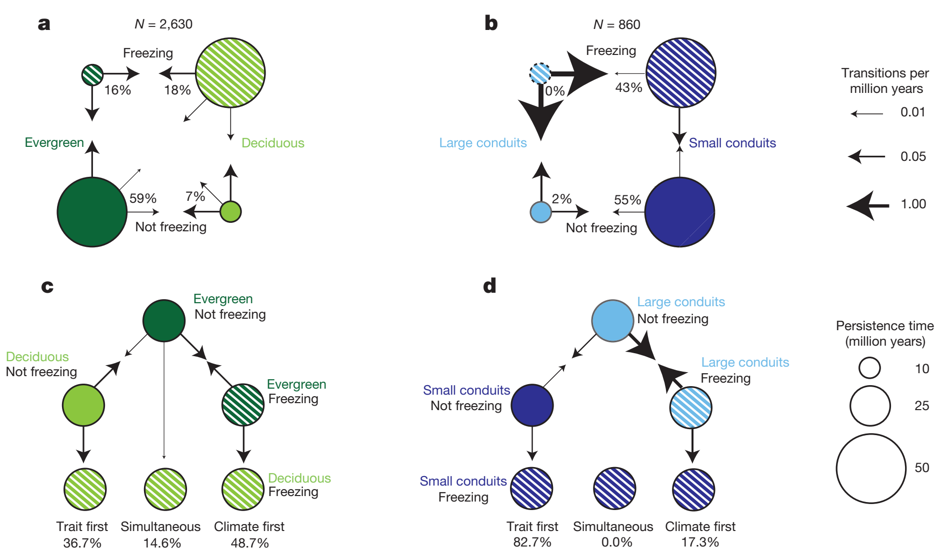


Figure 2 | Coordinated evolutionary transition rates between leaf phenology or conduit diameter and climate occupancy. a, b, A representation of coordinated evolution for the best likelihood-based model between leaf phenology for 2,630 species (evergreen, dark green; deciduous, light green) and climate occupancy (freezing exposed (freezing), striped; freezing unexposed (not freezing), solid) (**a**), and conduit diameter for 860 species (large (≥0.044 mm), light blue; small (<0.044 mm), dark blue) and climate occupancy (**b**) based on models fit to all Angiospermae. The sizes of the black arrows in the plot are proportional to the transition rates between each possible state combination (larger arrows denote higher rates; no arrows for rates of 0). The number at the top of each panel denotes the number of extant Angiospermae species used in the analyses and percentages denote the percentage of extant species with that character state. The size of each circle is proportional to the persistence time in that state, where persistence time is defined as the inverse of the sum of the transition rates away from a given character state (that is, the inverse of the sum of all arrow rates out of a character state). **c, d**, The relative likelihood of the different pathways out of the evergreen and freezing-unexposed state and into the deciduous and freezing-exposed state (**c**), and out of the large-diameter conduit and freezing-unexposed state and into the small-diameter conduit and freezing-exposed state (**d**). The three possible pathways between two focal character state combinations provide insight into whether lineages typically evolved: (1) with the trait first, such that phenology or conduit diameter shifted before encountering freezing; (2) with climate occupancy first, such that phenology or conduit diameter shifted after encountering freezing; or (3) both simultaneously, such that shifts in phenology or conduit diameter and encountering freezing happened at the same time (see Supplementary Information for further details).

and Superasteridae (superasterids) (see ref. 10 for lineage definitions); these clades represent ~ 22%, 3%, 34% and 34%, respectively, of all extant angiosperm species.

Across woody angiosperms, a model that assumed coordinated evolution of leaf phenology and climate occupancy was strongly supported over a model that assumed they evolved independently (Akaike information criteria (ΔAIC) = 310.1; Fig. 2a and Extended Data Table 2). Deciduous freezing-exposed and evergreen freezing-unexposed were highly persistent character states (Fig. 2a, as indicated by size of the circles, and Extended Data Table 3); persistence times (that is, expected time until state change) are defined as the inverse of the sum of estimated transition rates away from a given character state. Therefore, in the presence of freezing, the deciduous state was far more stable than the evergreen one. We also found that leaf phenology was generally about as labile as climate occupancy (climate:trait rate ratio = 0.845), and it was also far more likely to evolve as a response to a change in environment rather than arising before encountering freezing (that is, climate occupancy evolved first; Fig. 2c).

Similarly, across woody angiosperms, a model assuming coordinated evolution of conduit diameter size and climate occupancy was strongly supported over a model that assumed they evolved independently (ΔAIC = 21.5; Fig. 2b and Extended Data Table 2). Both climate occupancy states (freezing exposed and freezing unexposed) in combination with small conduits were highly persistent (Fig. 2b and Extended Data Table 3). Additionally, no species with large conduits were in the freezing-exposed state, indicating that this is a highly transitory character state (that is, short persistence time). As with leaf phenology, climate occupancy and conduit diameter were similar in their overall lability (climate:trait rate ratio = 0.895); however, a shift into environments with freezing temperatures was far more likely to occur after conduits had already shifted from large to small (that is, the trait evolved before climate occupancy; Fig. 2d).

Evolutionary shifts in growth habit were also strongly coordinated with shifts in climate. However, the nature of coordination varied considerably among major angiosperm clades (Extended Data Table 3), as did overall transition rates (superrosids and superasterids > magnoliids > monocots). Of 104 models evaluated, a 40-parameter model allowing each major lineage to have its own transition matrix received most support (Extended Data Table 4). These results were generally robust to uncertainty about whether species in the freezing-unexposed state actually lacked an ability to cope with freezing (Supplementary Information). Across angiosperms, asymmetry of transition rates led to numerous extant species in the woody freezing-unexposed and herbaceous freezing-exposed states (Fig. 3a and Extended Data Table 3). The large number of extant species in the woody freezing-unexposed state, according to our model, was the result of this state being persistent (Fig. 3a). Even within monocots, where relatively few woody species exist, the woody freezing-unexposed state was strongly persistent. The herbaceous freezing-exposed state, on the other hand, had low persistence times, indicating that the numerous extant species (N = 4,066 out of 12,706 species for which data are available) were due to many rapid transitions both into and out of this character state (Fig. 3a). Climate occupancy was much more labile than growth form (climate:trait rate ratio = 4.93). Furthermore, the predominant pathway within angiosperms from the woody freezing-unexposed state to the herbaceous freezing-exposed state was to first evolve the herbaceous habit and subsequently enter habitats with freezing-exposed conditions (that is, the trait evolved before the climate occupancy; Fig. 3b). This, in combination with the conduit diameter results, suggests that lineages that successfully colonized new freezing environments were probably predisposed to do so, at least for these two traits.

Although our focus here is on evolutionary links between species distributions with respect to freezing conditions and traits that allow species to cope with freezing, we note that differential diversification rates[23] and vagility among lineages also certainly played their parts in determining why we see species where we do today. For instance, herbs may have higher speciation and/or extinction rates than woody taxa[24]. Additionally, growth form may influence a plant's ability to disperse to and colonize newly emerging locations with freezing temperatures[25]. Tests of these alternatives are critical for fully understanding how angiosperms radiated into freezing environments, but such analyses require an even more complete record of global distributions of vagility and growth habit across land plants and a comparably more completely sampled phylogeny. These are non-trivial improvements as we currently have growth habit data for only 16% of accepted land plants[20] (R.G.F. *et al.*, manuscript submitted) and molecular and climate data for 26% (12,706 species) of those taxa. Total trait records are fewer for phenology (6,705 species) and conduit diameter (2,181 species).

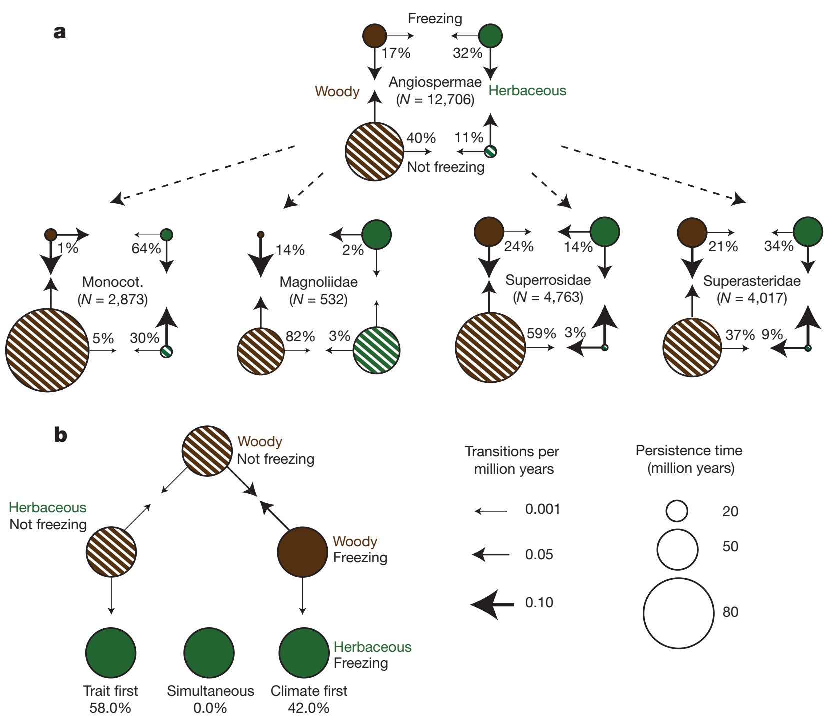


Figure 3 | Coordinated evolutionary transition rates between growth form and climate occupancy. **a**, A representation of coordinated evolution for the best likelihood-based model between growth form for 12,706 species (herbaceous, green; woody, brown) and climate occupancy based on a model assuming the same rates were applied to all Angiospermae (top plot above the dashed arrow), and the best-fit model, in which rates were estimated separately for the major lineages, that is, Monocotyledoneae, Magnoliidae, Superrosidae and Superasteridae (bottom four plots below the dashed arrows). **b**, The weighted average (by clade diversity) of the relative likelihood of the different pathways out of the woody and freezing-unexposed state and into the herbaceous and freezing-exposed state (see Fig. 2 and Methods for further details).

Among three key angiosperm strategies successful in today's freezing environments (deciduous leaves, small conduits and herbaceous habit), our analyses indicated two especially striking findings. First, the pathway to herbaceousness or small conduits in freezing environments largely involved acquisition of the trait first (followed by adaptation to a new climate), whereas the pathway to deciduousness in freezing environments was largely via a shift in climate occupancy first (followed by evolution of the trait). Second, transitions between growth habit states should be fairly simple genetically[26], involving suppression and re-expression of only a few genes[27], and, traditionally, growth habit has been considered highly labile (ref. 6, but see refs 16, 28, 29). Our results are consistent with climate occupancy being more labile than growth habit, and freezing environments being largely filled by a subset of lineages that were already herbaceous or, if woody, had small conduits before they encountered freezing. Why these lineages initially evolved a herbaceous habit and small conduit sizes remains unclear; these traits are probably tightly associated with responses to other environmental gradients (for example, aridity in the tropics) and numerous other aspects of a plant's ecological strategy (for example, seed size, tissue defence, and so on) related to resource acquisition and disturbance regimes. Therefore, successful shifts between stem constructions take more than just turning on or off a few genes.

By weaving together a series of disparate threads encapsulating evolution, functional ecology and the biogeographic history of angiosperms, including extensive functional trait databases and an exceptionally large timetree, we have documented the likely evolutionary pathways of trait acquisition facilitating angiosperm radiation into the cold.

METHODS SUMMARY

To examine the evolutionary responses to freezing in angiosperms, we first compiled trait data on leaves and stems from existing databases and the literature. Growth form data came from numerous sources and were coded as a binary trait (woody or herbaceous; Supplementary Table 1). Leaf phenology and conduit diameter came from existing databases (see Supplementary Information for a list). Second, taxonomic nomenclature was made consistent among data sets and up to date by querying species names against the International Plant Names Index (http://www.ipni.org/), Tropicos (http://www.tropicos.org/), The Plant List (http://www.theplantlist.org/) and the Angiosperm Phylogeny website (http://www.mobot.org/MOBOT/research/APweb/). Third, we obtained species' spatial distributions from Global Biodiversity Information Facility records (http://www.gbif.org/; Supplementary Table 4) and then determined whether species encountered freezing temperatures using climate data from the WorldClim database (http://www.worldclim.org/). Fourth, we constructed a dated phylogeny for these species by downloading available GenBank sequences (http://www.ncbi.nlm.nih.gov/genbank/) for seven gene regions. Genetic data were compiled and aligned using the PHLAWD pipeline (v.3.3a), and maximum-likelihood-based phylogenetic analyses of the total sequence alignment were performed using RAxML (v.7.4.1), partitioned by gene region and with major clades (that is, families and orders) constrained according to the APG III classification system. Branch lengths were time-scaled using congruification, which involved using divergence times estimated from a reanalysis of a broadly sampled data set (Extended Data Fig. 2 and Supplementary Tables 2 and 3). Last, tests of coordinated evolution among traits in our database were analysed in the corHMM R package; transition rates between two binary traits were analysed using a likelihood-based model.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 July; accepted 5 November 2013.

Published online 22 December 2013; corrected online 3 January 2014 (see full-text HTML version for details).

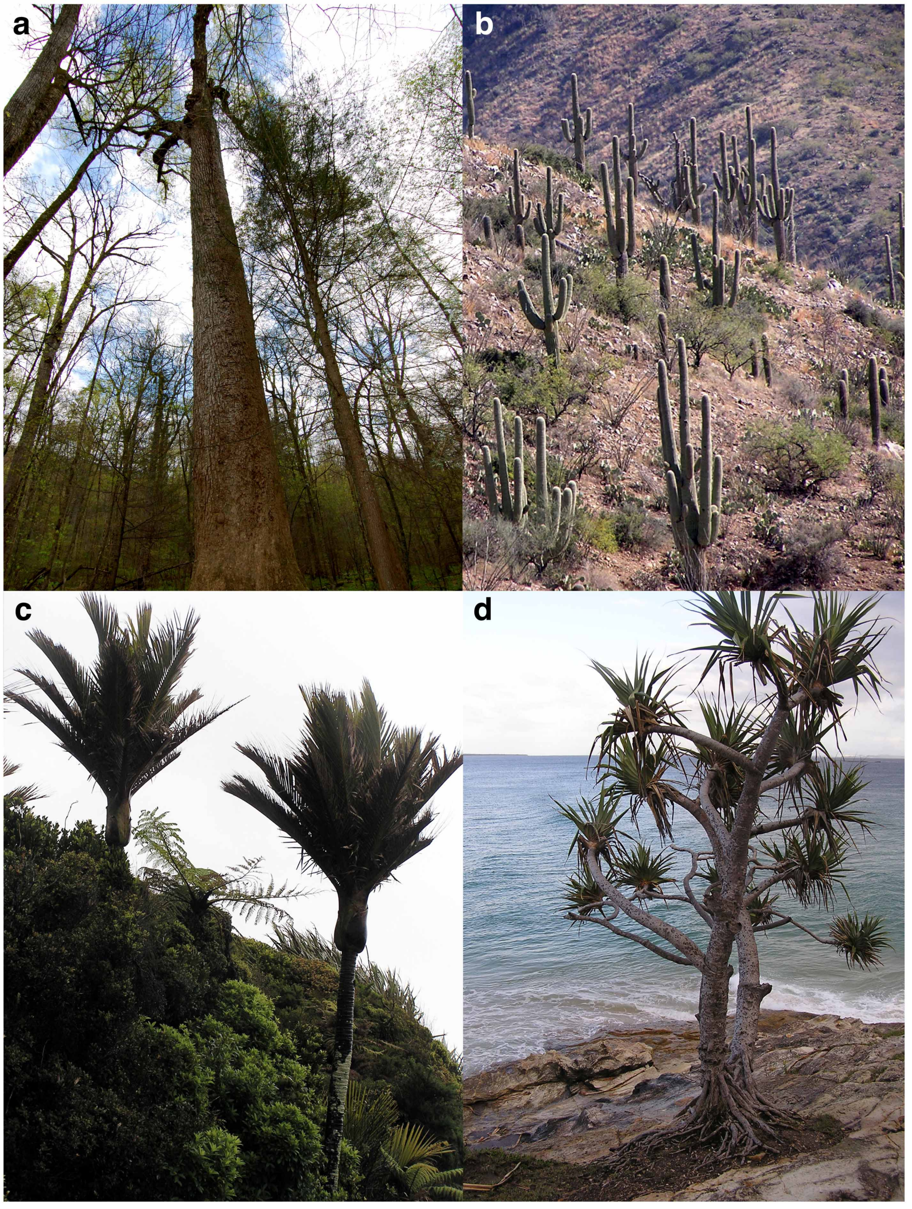
1. Sinnott, E. W. & Bailey, I. W. The evolution of herbaceous plants and its bearing on certain problems of geology and climatology. *J. Geol.* **23,** 289–306 (1915).
2. Wing, S. L. & Boucher, L. D. Ecological aspects of the Cretaceous flowering plant radiation. *Annu. Rev. Earth Planet. Sci.* **26,** 379–421 (1998).
3. Feild, T. S., Arens, N. C., Doyle, J. A., Dawson, T. E. & Donoghue, M. J. Dark and disturbed: a new image of early angiosperm ecology. *Paleobiology* **30,** 82–107 (2004).
4. Moles, A. T. *et al.* Global patterns in plant height. *J. Ecol.* **97,** 923–932 (2009).
5. Tyree, M. T. & Zimmermann, M. H. *Xylem Structure and the Ascent of Sap* (Springer, 2002).
6. Cronquist, A. *The Evolution and Classification of Flowering Plants*. (Houghton Mifflin, 1968).
7. Kattge, J. *et al.* TRY—a global database of plant traits. *Glob. Change Biol.* **17,** 2905–2935 (2011).
8. Stebbins, G. L. The probable growth habit of the earliest flowering plants. *Ann. Mo. Bot. Gard.* **52,** 457–468 (1965).
9. Taylor, D. & Hickey, L. Phylogenetic evidence for the herbaceous origin of angiosperms. *Plant Syst. Evol.* **180,** 137–156 (1992).
10. Soltis, D. E. *et al.* Angiosperm phylogeny: 17 genes, 640 taxa. *Am. J. Bot.* **98,** 704–730 (2011).
11. Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc. Natl Acad. Sci. USA* **107,** 5897–5902 (2010).
12. Spicer, R. & Groover, A. Evolution of development of vascular cambia and secondary growth. *New Phytol.* **186,** 577–592 (2010).
13. Feild, T. S. & Wilson, J. P. Evolutionary voyage of angiosperm vessel structure-function and its significance for early angiosperm success. *Int. J. Plant Sci.* **173,** 596–609 (2012).
14. Philippe, M. *et al.* Woody or not woody? Evidence for early angiosperm habit from the Early Cretaceous fossil wood record of Europe. *Palaeoworld* **17,** 142–152 (2008).
15. Wiens, J. J. & Donoghue, M. J. Historical biogeography, ecology and species richness. *Trends Ecol. Evol.* **19,** 639–644 (2004).
16. Donoghue, M. J. A phylogenetic perspective on the distribution of plant diversity. *Proc. Natl Acad. Sci. USA* **105,** 11549–11555 (2008).
17. Wheeler, E. A., Baas, P. & Rodgers, S. Variations in dicot wood anatomy: a global analysis based on the Insidewood database. *IAWA J.* **28,** 229–258 (2007).
18. Botta, A., Viovy, N., Ciais, P., Friedlingstein, P. & Monfray, P. A global prognostic scheme of leaf onset using satellite data. *Glob. Change Biol.* **6,** 709–725 (2000).
19. Judd, W. S., Sanders, R. W. & Donoghue, M. J. Angiosperm family pairs: preliminary phylogenetic analysis. *Harv. Pap. Bot.* **5,** 1–49 (1994).
20. Paton, A. J. *et al.* Towards target 1 of the global strategy for plant conservation: a working list of all known plant speciesprogress and prospects. *Taxon* **57,** 602–611 (2008).
21. Loehle, C. Height growth rate tradeoffs determine northern and southern range limits for trees. *J. Biogeogr.* **25,** 735–742 (1998).
22. Davis, S. D., Sperry, J. S. & Hacke, U. G. The relationship between xylem conduit diameter and cavitation caused by freezing. *Am. J. Bot.* **86,** 1367–1372 (1999).
23. Maddison, W. P. Confounding asymmetries in evolutionary diversification and character change. *Evolution* **60,** 1743–1746 (2006).
24. Soltis, D. E. *et al.* Phylogenetic relationships and character evolution analysis of Saxifragales using a supermatrix approach. *Am. J. Bot.* **100,** 916–929 (2013).
25. Thomson, F. J., Moles, A. T., Auld, T. D. & Kingsford, R. T. Seed dispersal distance is more strongly correlated with plant height than with seed mass. *J. Ecol.* **99,** 1299–1307 (2011).
26. Groover, A. T. What genes make a tree a tree? *Trends Plant Sci.* **10,** 210–214 (2005).
27. Lens, F., Smets, E. & Melzer, S. Stem anatomy supports *Arabidopsis thaliana* as a model for insular woodiness. *New Phytol.* **193,** 12–17 (2012).
28. Jansson, R., Rodríguez-Castañeda, G. & Harding, L. E. What can multiple phylogenies say about the latitudinal diversity gradient? A new look at the tropical conservatism, out-of-the-tropics and diversification rate hypotheses. *Evolution* **67,** 1741–1755 (2013).
29. Beaulieu, J. M., O'Meara, B. C. & Donoghue, M. J. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Syst. Biol.* **62,** 725–737 (2013).

Supplementary Information is available in the online version of the paper.

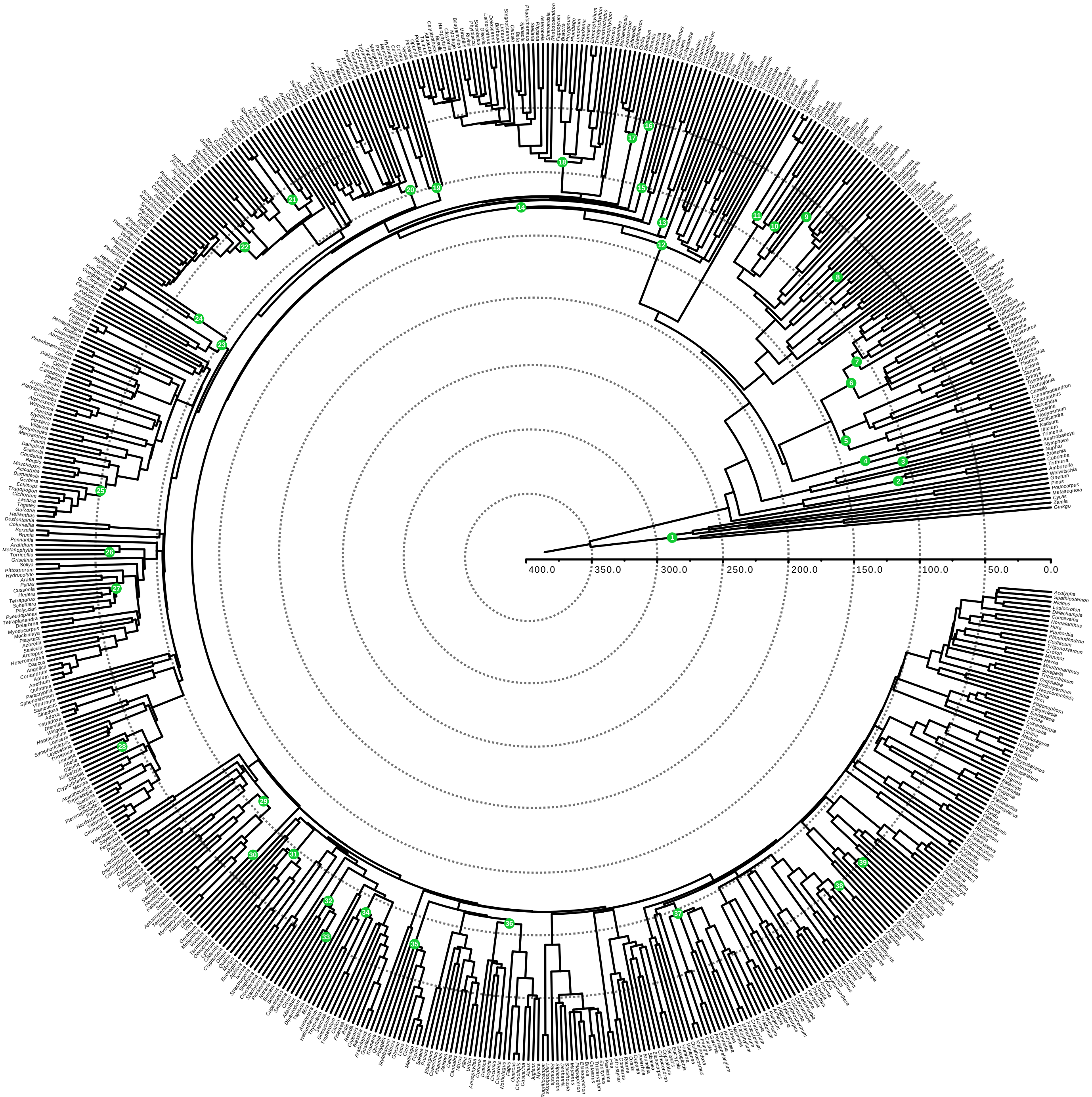
Acknowledgements We thank T. Robertson and A. Hahn at the Global Biodiversity Information Facility for providing species' georeference points, A. Ordonez for providing growth form data, and A. Miller and D. Ackerly for helpful comments on a draft of this manuscript. Support for this work was given to the working group "Tempo and Mode of Plant Trait Evolution: Synthesizing Data from Extant and Extinct Taxa" by the National Evolutionary Synthesis Center (NESCent), National Science Foundation grant #EF- 0905606 and Macquarie University Genes to Geoscience Research Centre.

Author Contributions A.E.Z., W.K.C., D.C.T. and J.M.B. designed the initial project, wrote the original manuscript and carried out analyses. J.M.E., S.A.S. and D.C.T. constructed the timetree. J.M.E., R.G.F., D.J.M., B.C.O'M. and S.A.S. were major quantitative contributors, especially with the development of new methods, analyses, graphics and writing. A.T.M., P.B.R., D.L.R., D.E.S., P.F.S., I.J.W. and M.W. were large contributors through the development of initial ideas, methods, dataset curation, analyses and writing. L.A., R.I.B., A.C., R.G., F.H., M.R.L., J.O., P.S.S., N.G.S. and L.W. contributed data sets and discussions, and read drafts.

Author Information Data and code are deposited at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.63q27) and TRY (http://www.try-db.org/). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.E.Z. (aezanne@gmail.com).



Extended Data Figure 1 | Examples of the definition of 'woody'. a–d, We defined 'woody' as having a prominent aboveground stem that is persistent over time and with changing environmental conditions. **a**, *Liriodendron tulipifera* (Magnoliaceae), Joyce Kilmer Memorial Forest, Robbinsville, North Carolina, USA. **b**, *Carnegiea giganteana* (Cactaceae), Biosphere II, Tucson, Arizona, USA, **c**, *Rhopalostylis sapida* (Arecaceae) and *Cyathea sp.* (Cyatheaceae), Punakaiki, South Island, New Zealand. **d**, *Pandanus sp.* (Pandanaceae), Moreton Bay Research Station, North Stradbroke Island, Queensland, Australia (photographs by A.E.Z.).



Extended Data Figure 2 | Reference timetree used for congruification analyses. Results of the divergence time estimation of 639 taxa of seed plants from the reanalysis of a previously described[10] phylogeny. Fossil calibrations are indicated at the nodes with green circles, and numbers correspond to fossils described in Supplementary Table 2. Concentric dashed circles represent 100-Myr intervals as indicated by the scale bar.

Extended Data Table 1 | Number of species in different growth forms by clade

Lineage	Woody	Herbaceous	Total	Proportion herbaceous
Angiospermae	28650	17347	45997	0.38
Magnoliidae	2438	75	2513	0.03
Monocotyledoneae	1226	9894	11120	**0.89**
Superasteridae	8468	4863	13331	0.36
Superrosidae	14885	1956	16841	0.12
ANA grade+Chloranthales				
Amborellales	1	0	1	0.00
Austrobaileyales	48	0	48	0.00
Chloranthales	18	7	25	0.28
Nymphaeales	0	43	43	**1.00**
Magnoliidae				
Canellales	71	0	71	0.00
Laurales	1212	6	1218	0.00
Magnoliales	1053	0	1053	0.00
Piperales	102	69	171	0.40
Monocotyledoneae				
Acorales	0	7	7	**1.00**
Alismatales	3	513	516	**0.99**
Arecales	793	0	793	0.00
Asparagales	141	4133	4274	**0.97**
Commelinales	0	180	180	**1.00**
Dioscoreales	0	178	178	**1.00**
Liliales	35	459	494	**0.93**
Pandanales	80	17	97	0.18
Petrosaviales	0	3	3	**1.00**
Poales	109	4075	4184	**0.97**
Zingiberales	61	329	390	**0.84**
Basal eudicots+Gunnerales				
Buxales	31	0	31	0.00
Ceratophyllales	0	3	3	**1.00**
Gunnerales	2	14	16	**0.88**
Proteales	1354	3	1357	0.00
Ranunculales	134	488	622	**0.78**
Trochodendrales	2	0	2	0.00
Superasteridae				
Apiales	410	226	636	0.36
Aquifoliales	211	0	211	0.00
Asterales	548	1775	2323	**0.76**
Berberidopsidales	3	0	3	0.00
Bruniales	65	0	65	0.00
Caryophyllales	545	712	1257	**0.57**
Cornales	163	68	231	0.29
Dilleniales	71	0	71	0.00
Dipsacales	151	61	212	0.29
Ericales	2798	350	3148	0.11
Escalloniales	23	0	23	0.00
Garryales	17	0	17	0.00
Gentianales	1508	280	1788	0.16
Lamiales	1214	1035	2249	0.46
Paracryphiales	20	0	20	0.00
Santalales	242	20	262	0.08
Solanales	254	200	454	0.44
Superrosidae				
Brassicales	136	389	525	**0.74**
Celastrales	228	11	239	0.05
Crossosomatales	31	0	31	0.00
Cucurbitales	62	169	231	**0.73**
Fabales	2462	448	2910	0.15
Fagales	745	0	745	0.00
Geraniales	27	63	90	**0.70**
Huerteales	8	0	8	0.00
Malpighiales	2978	294	3272	0.09
Malvales	1195	64	1259	0.05
Myrtales	2787	79	2866	0.03
Oxalidales	396	14	410	0.03
Picramniales	16	0	16	0.00
Rosales	1465	143	1608	0.09
Sapindales	2082	7	2089	0.00
Saxifragales	190	246	436	**0.56**
Vitales	42	1	43	0.02
Zygophyllales	35	12	47	0.26

Number of species that are woody, number of species that are herbaceous, total number of species, and proportion of herbaceous species in major lineages and orders. Proportions in bold are lineages with >0.5 species that are herbaceous.

Extended Data Table 2 | Coordinated evolutionary model fits for leaf phenology, conduit diameter and climate occupancy

Leaf Phenology and climate occupancy					
Model	Number of parameters	-lnL	AIC	ΔAIC	w_i
Character independent	4	-2305.4	4618.9	312.8	<0.01
Character dependent, equal rates	1	-2401.3	4804.5	498.4	<0.01
Character dependent, all rates diff	8	-2160.0	4336.0	29.9	<0.01
Character dependent, all rates diff*	***12***	***-2141.1***	***4306.1***	***0***	***0.99***
Conduit diameter and climate occupancy					
Model	Number of parameters	-lnL	AIC	ΔAIC	w_i
Character independent	4	-603.65	1223.3	21.5	<0.01
Character dependent, equal rates	1	-739.8	1481.6	279.8	<0.01
Character dependent, all rates diff	***8***	***-592.91***	***1201.8***	***0***	***0.98***
Character dependent, all rates diff*	12	-592.91	1209.8	8.0	0.02

The likelihood-based best model in each case (shown in bold italics) was chosen based on both AIC and Akaike weights (w_i). Also listed for each model are the number of parameters, negative log likelihood (−lnL), and ΔAIC. The asterisk indicates a model where simultaneous changes in any two binary characters were allowed to change.

Extended Data Table 3 | Coordinated evolutionary model transition rates

Leaf Phenology and climate occupancy		Conduit Diameter and climate occupancy	
Transition	Angiospermae transition rates	Transition	Angiospermae transition rates
EVERGREEN EXPOSED→EVERGREEN UNEXPOSED	0.051 (0.042,0.065)	LARGE EXPOSED→LARGE UNEXPOSED	100.0 (0.000,100.0)
DECIDUOUS UNEXPOSED→EVERGREEN UNEXPOSED	0.053 (0.053,0.097)	SMALL UNEXPOSED→LARGE UNEXPOSED	0.005 (0.003,0.041)
DECIDUOUS EXPOSED→EVERGREEN UNEXPOSED	0.005 (0.004,0.006)	SMALL EXPOSED→LARGE UNEXPOSED	0.000 (na,na)
EVERGREEN UNEXPOSED→EVERGREEN EXPOSED	0.011 (0.001,0.014)	LARGE UNEXPOSED→LARGE EXPOSED	0.033 (0.000,0.190)
DECIDUOUS UNEXPOSED→EVERGREEN EXPOSED	0.0023 (0.000,0.003)	SMALL UNEXPOSED→LARGE EXPOSED	0.000 (na,na)
DECIDUOUS EXPOSED→EVERGREEN EXPOSED	0.018 (0.012,0.019)	SMALL EXPOSED→LARGE EXPOSED	0.000 (0.000,0.000)
EVERGREEN UNEXPOSED→DECIDUOUS UNEXPOSED	0.008 (0.008,0.012)	LARGE UNEXPOSED→SMALL UNEXPOSED	0.096 (0.065,1.07)
EVERGREEN EXPOSED→DECIDUOUS UNEXPOSED	0.0000 (0.000,0.001)	LARGE EXPOSED→SMALL UNEXPOSED	0.000 (na,na)
DECIDUOUS EXPOSED→DECIDUOUS UNEXPOSED	0.002 (0.001,0.002)	SMALL EXPOSED→SMALL UNEXPOSED	0.0353 (0.026,0.038)
EVERGREEN UNEXPOSED→DECIDUOUS EXPOSED	0.001 (0.000,0.001)	LARGE UNEXPOSED→SMALL EXPOSED	0.000 (na,na)
EVERGREEN EXPOSED→DECIDUOUS EXPOSED	0.0116 (0.009,0.014)	LARGE EXPOSED→SMALL EXPOSED	100.00 (0.000,100.0)
DECIDUOUS UNEXPOSED→DECIDUOUS EXPOSED	0.0116 (0.010,0.019)	SMALL UNEXPOSED→SMALL EXPOSED	0.0225 (0.017,0.026)

Growth habit and climate occupancy					
Transition	Monocotyledonae transition rates	Magnoliidae transition rates	Superrosidae transition rates	Superasteridae transition rates	Rest transition rates
WOODY EXPOSED→WOODY UNEXPOSED	0.044 (0.05,0.159)	0.126 (0.045,0.112)	0.030 (0.027,0.035)	0.041 (0.031,0.049)	0.021 (0.007,0.020)
HERBACEOUS UNEXPOSED→WOODY UNEXPOSED	0.001 (0.000,0.001)	0.002 (0.000,0.010)	0.049 (0.041,0.065)	0.052 (0.055,0.076)	0.000 0.000,0.003)
HERBACEOUS EXPOSED→WOODY UNEXPOSED	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)
WOODY UNEXPOSED→WOODY EXPOSED	0.005 (0.008,0.027)	0.017 (0.008,0.019)	0.001 (0.009,0.012)	0.0189 (0.016,0.024)	0.028 (0.016,0.031)
HERBACEOUS UNEXPOSED→WOODY EXPOSED	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)
HERBACEOUS EXPOSED→WOODY EXPOSED	0.001 (<0.001,0.001)	0.016 (0.001,0.021)	0.008 (0.007,0.009)	0.012 (0.011,0.013)	0.001 (<0.001,0.003)
WOODY UNEXPOSED→HERBACEOUS UNEXPOSED	0.001 (0.000,0.001)	0.001 (<0.001,0.001)	0.002 (0.001,0.002)	0.004 (0.002,0.005)	<0.001 (0.000,<0.001)
WOODY EXPOSED→HERBACEOUS UNEXPOSED	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)
HERBACEOUS EXPOSED→HERBACEOUS UNEXPOSED	0.0483 (0.037,0.086)	0.003 (<0.001,0.036)	0.024 (0.017,0.036)	0.045 (0.028,0.062)	0.003 (0.003,0.022)
WOODY UNEXPOSED→HERBACEOUS EXPOSED	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)	0.000 (na,na)
WOODY EXPOSED→HERBACEOUS EXPOSED	0.007 (0.002,0.019)	0.000 (0.000,0.003)	0.002 (0.001,0.005)	0.004 (0.002,0.005)	0.003 (0.002,0.004)
HERBACEOUS UNEXPOSED→HERBACEOUS EXPOSED	0.060 (0.056,0.129)	0.015 (0.011,0.042)	0.090 (0.050,0.139)	0.147 (0.101,0.232)	0.033 (0.031,0.304)

The estimated transition rates for the best likelihood-based evolutionary transitions model between climate occupancy and either growth habit, leaf phenology or conduit diameter evolution are included. The numbers in parentheses denote the values at the 2.5% and 97.5% quantiles of the distribution of parameter estimates obtained from the same analyses run on the 100 bootstrapped trees (see Supplementary Information). The leaf phenology model includes transitions between combinations of leaf phenology (evergreen, deciduous) and climate occupancy (freezing exposed, freezing unexposed), the conduit diameter model includes transitions between combinations of conduit diameter (large ≥0.044 mm, small <0.044 mm) and climate occupancy, and the growth habit model includes transitions between combinations of growth form (herbaceous, woody) and climate occupancy. Arrows denote the direction of the transition. The growth habit model assumes separate models for the major groups within angiosperms: Monocotyledonae, Magnoliidae, Superrosidae, Superasteridae and all remaining angiosperms (the rest), including the ANA grade, Chloranthales, Ceratophyllales and basal eudicots plus Gunnerales. The leaf phenology and conduit diameter models assume a single model for all angiosperms.

Extended Data Table 4 | Coordinated evolutionary model fits for growth form and climate occupancy

Model	Number of parameters	-lnL	AIC	ΔAIC	w_i
ABCDE	40	-8348.9	16777.9	0	0.999
AABCD	48*	-8347.7	16791.3	13.4	<0.001
AABCD	32	-8353.9	16794.4	16.5	<0.001

The top three of 104 likelihood-based models tested for growth form and climate occupancy evolution are reported. The best model, based on both AIC and Akaike weights (w_i), was a model that assigned a separate rate for the Monocotyledonae (position 1), Magnoliidae (position 2), Superrosidae (position 3), Superasteridae (position 4) and all remaining angiosperms, including the ANA grade, Chloranthales, Ceratophyllales and basal eudicots plus Gunnerales (position 5), respectively. Also listed for each model are the number of parameters, negative log likelihood (−lnL), and ΔAIC. The asterisk indicates a model where simultaneous changes in any two binary characters were allowed.

doi:10.1038/nature12786

Modelling the effects of subjective and objective decision making in scientific peer review

In-Uck Park[1,2], Mike W. Peacey[1,3] & Marcus R. Munafò[4,5,6]

The objective of science is to advance knowledge, primarily in two interlinked ways: circulating ideas, and defending or criticizing the ideas of others. Peer review acts as the gatekeeper to these mechanisms. Given the increasing concern surrounding the reproducibility of much published research[1], it is critical to understand whether peer review is intrinsically susceptible to failure, or whether other extrinsic factors are responsible that distort scientists' decisions. Here we show that even when scientists are motivated to promote the truth, their behaviour may be influenced, and even dominated, by information gleaned from their peers' behaviour, rather than by their personal dispositions. This phenomenon, known as herding, subjects the scientific community to an inherent risk of converging on an incorrect answer and raises the possibility that, under certain conditions, science may not be self-correcting. We further demonstrate that exercising some subjectivity in reviewer decisions, which serves to curb the herding process, can be beneficial for the scientific community in processing available information to estimate truth more accurately. By examining the impact of different models of reviewer decisions on the dynamic process of publication, and thereby on eventual aggregation of knowledge, we provide a new perspective on the ongoing discussion of how the peer-review process may be improved.

Current incentive structures in science promote attempts to publish in prestigious journals, which frequently prioritize new, exciting findings. One consequence of this may be the emergence of fads and fashions in the scientific literature (that is, 'hot topics')[1], leading to convergence on a particular paradigm or methodology. This may not matter if this convergence is on the truth—topics may simply cease to be hot topics as the problem becomes solved. However, there is increasing concern that many published research findings are in fact false[1]. It is common for early findings to be refuted by subsequent evidence, often leading to the formation of groups that interpret the same evidence in notably different ways[2], and this phenomenon is observed across many scientific disciplines[3,4]. There are a number of relatively recent examples of convergence on false hypotheses, such as the theory of stress causing gastric ulcer formation[5]. Once established, these can become surprisingly difficult to refute[6]—they may become "more 'vampirical' than 'empirical'—unable to be killed by mere evidence"[7]. Science may therefore not be as self-correcting as is commonly believed[8], and the selective reporting of results can produce literatures that "consist in substantial part of false conclusions"[9].

It is important to understand how convergence on false conclusions may come about. A number of possibilities present themselves. First, scientists may not in fact be rational individuals pursuing the truth after all—an argument made by some influential sociologists of science (the strong programme)[10]—or may be rational but stuck within a particular paradigm[11]. Second, some scientists may be biased or even immoral—a number of high profile cases of data fabrication and fraud have emerged in recent years[12]. Third, some scientists may care more about publication and careers than discovering the truth (that is, 'publish or perish'), a process which may be conscious or unconscious[13]. In competitive fields current incentive structures prioritize positive results, which may increase the likelihood of modification of data or conducting many statistical tests to achieve these; similarly, increased error rates may arise from multiple competing research groups testing the same hypotheses[14].

It has been shown that increased popularity of a particular research theme reduces the reliability of published results[14], and that findings published in prestigious journals are less reliable and more likely to be retracted[15]. Therefore, the convergence of research interest on a current hot topic may serve to undermine the reliability and veracity of subsequently published findings. In principle, peer review should eliminate or reduce these problems but, given empirical evidence for the unreliability of much published research, it may not in fact be conducted properly, or the process itself may be flawed. Empirical research and simulations have identified a number of factors which contribute to the likelihood that a published research finding is false[1,16]. However, the peer-review process itself has not been closely investigated as a possible influence, despite the fact that it acts as the ultimate gatekeeper of research publication. It is generally regarded as imperfect, although still the best model available to ensure both the quality and veracity of published scientific research, but there has been growing concern that it fails, at least in part, with respect to each of these two goals[1].

To understand the peer-review mechanism better, using a Bayesian approach in a model of the publication process, we analysed the behaviour of scientists who have developed their initial opinions independently as to which of the two opposing hypotheses, A and B, is more likely to be true. They know that on average their opinion is indeed correct with probability $\beta \in \left(\frac{1}{2}, 1\right)$, so they feel confident, but less than fully, about their opinion. The more controversial the issue, the lower the value of β. Upon receiving a manuscript that advocates one of the hypotheses, the editor of a hypothetical journal solicits a review from another scientist, who recommends acceptance or rejection. To focus on the influence of reviewer behaviour, rather than that of editor, we assume that the editor simply follows the reviewer's recommendation. Subsequently, the reviewer writes and submits their own manuscript to the journal, and the process repeats. The two decisions for each scientist are therefore: (1) whether or not to recommend acceptance of a manuscript that they are reviewing, and (2) which hypothesis to advocate in their own submission, which we term the 'theme' of their manuscript. As a publication history evolves (following cycles of submission, peer review and acceptance or rejection) a scientist revises their view on the likelihood of each hypothesis being true, in light of the relative probability of this particular history occurring when one hypothesis is true as opposed to the other. Being motivated to promote the truth, each scientist will advocate a theme that is more likely to be true, according to their revised view when they submit a manuscript.

Our aim was to understand how different criteria of reviewing decisions influence the publication outcome, and how the resulting publication

[1]Department of Economics, University of Bristol, Bristol BS8 1TN, UK. [2]Department of Economics, Sungkyunkwan University, Seoul 110-745, South Korea. [3]Department of Economics, University of Bath, Bath BA2 7AY, UK. [4]MRC Integrative Epidemiology Unit (IEU), University of Bristol, Bristol BS8 1BN, UK. [5]UK Centre for Tobacco and Alcohol Studies, University of Bristol, Bristol BS8 1TU, UK. [6]School of Experimental Psychology, University of Bristol, Bristol BS8 1TU, UK.

histories and the information inherent in the relevant peer-review criterion influence the community's eventual understanding of the topic. To this end we modelled and compared two different ways that scientists approach the reviewing decision. In the first model (M1), the subjective criterion of how strongly the reviewer agrees with the conclusion of the research (that is, the theme of the manuscript) is reflected in the decision, in addition to other more objective criteria such as research design and methodology. In the second model (M2), the decision reflects objective criteria only. Our findings, therefore, may shed light on whether subjective assessment is desirable in the peer-review process and, if so, to what extent. As a benchmark, we also compared M1 and M2 with a default model (M3), in which all manuscripts are published without any filtering through peer review. As scientists will make inferences that take into account how reviewers arrive at their recommendations, the particular peer-review model in operation affects how they revise their views and, thereby, their decisions on which theme to advocate as an author, as well as their decisions as a reviewer.

The results of the three models (Fig. 1) indicate that: (1) almost certainly, some scientists will submit manuscripts on themes which disagree with their initial opinion (we term this 'herding'); (2) the extent to which the wider scientific community's perception of a literature is removed from the truth (we term this 'misperception') decreases with number of publications, but information transmission is greatly hampered once herding has occurred, to such an extent that no further improvement in understanding occurs except in M1 where a degree of subjectivity is allowed in the reviewing decision (that is, reviewers as well as authors act guided by Bayesian inference); and (3) the probability of another publication on a particular issue increases as the number of manuscripts published on that issue increases, owing to aggregation of information and herding reinforcing the scientific community's consensus.

The phenomenon known as herding is inherent in the behaviour of scientists operating under all of the models we consider. An individual is said to be herding if they choose a theme to advocate in their manuscript submission based entirely on what they have observed from others, independently of what they initially thought was true. The degree of herding depends on the peer-review model in operation, the number of manuscripts submitted so far, and how confident scientists feel about their initial opinion (β). Herding takes place relatively quickly (Fig. 1), and we observe discrete jumps in the measure of herding early on in the process, when each signal (that is, the information carried by a peer-review decision) carries a large weighting. Notably, the probability of herding and the speed with which it increases are eventually lower when a degree of subjectivity is allowed in the reviewing decision (M1), and only in this case can a fad be reversed following a sequence of publications on the same theme. As a fad persists, the total number of scientists required in order to reverse this fad increases—and at a faster rate.

We use 'misperception' to describe how incorrect the perception of the wider scientific community is after a history of publication outcomes. It is defined as the probability that an outsider assigns to a hypothesis being correct, based on Bayesian inference from the observed history, when it is actually incorrect. The level of expected misperception (Fig. 1) remains relatively stable for low and high values of β, but for intermediate values of β it declines with increasing numbers of submitted manuscripts. Critically, when a degree of subjectivity is allowed in the peer-review process (M1), this always eventually outperforms the other models, because in these models information completely fails to be transmitted after herding occurs.

In our models, manuscript submission decisions made by individual scientists are based in part on information inferred from others' actions, because individuals use information from the publication history within a particular field, as well as their personal opinions, to guide their decisions. This may have positive effects if the decisions cluster around a correct outcome, or have negative effects if they cluster around an incorrect outcome. A degree of subjectivity in the peer-review process will, on average, lead to lower misperception, because reviewer decisions (and subsequent editorial decisions) which go against the herding trend will continue to reveal new information. In addition, the process is dynamic, and we show that self-correction can eventually occur when a degree of subjectivity is allowed in the peer-review process; however, it may not when the reviewing decision is completely independent of the reviewer's subjective assessment of the theme of the manuscript, and is based only on other, largely objective characteristics of the manuscript, such as the quality of the research methodology. In this case the probability of herding reaches 1 within finite time for all values of β, and the level of misperception cannot go below a certain lower bound. The concept of herding has been discussed in the context of scientific research in the past[17], but ours is the first study, to our knowledge, to model the processes by which it may occur.

These results raise the question of whether a higher level of subjectivity in reviewer decisions will lead to more effective restraint of incorrect herding. We therefore decided to test generalized M1 models, in which we varied the degree to which the reviewer's recommendation is determined by their subjective assessment of the conclusion. Our results (Fig. 2) indicate that excessively subjective reviews are not effective in restraining incorrect herding. This is because, in this case, recommendations are sensitive to whether the conclusion agrees with the reviewer's viewpoint at that time, and this factor is predominantly determined by

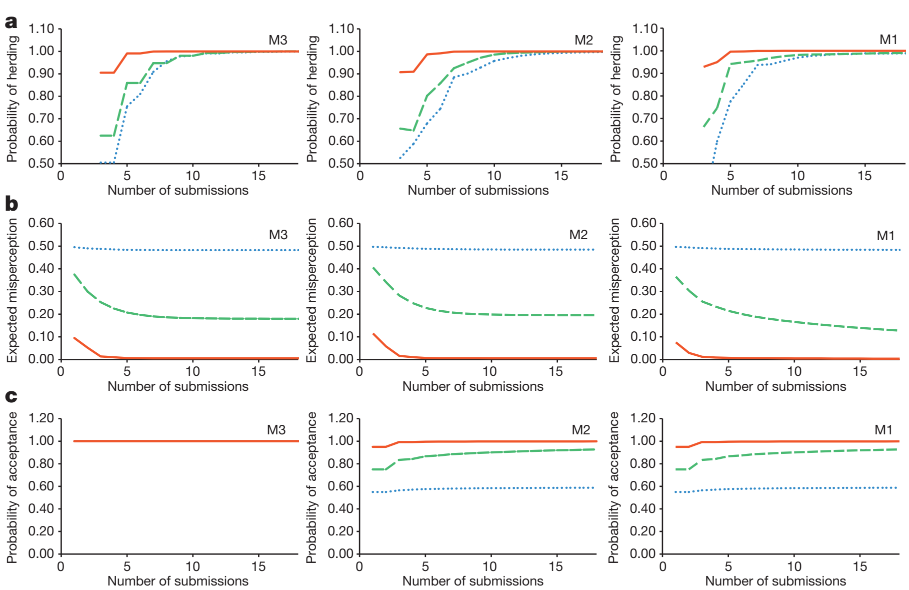


Figure 1 | Three models of peer-review behaviour. We show three models, M1 (right), M2 (middle) and M3 (left), which differ in the extent to which the peer-review decision depends on whether the reviewer agrees with the conclusion. Three outcomes are presented: (1) probability of herding (top), (2) average misperception generated (middle), and (3) probability of acceptance (bottom). The probability that the initial opinion is correct is reflected by β, and each outcome is presented for three values of β: (1) 0.55 (blue, dotted line), (2), 0.75 (green, dashed line), and (3) 0.95 (red, solid line), reflecting high, intermediate, and low uncertainty, respectively.

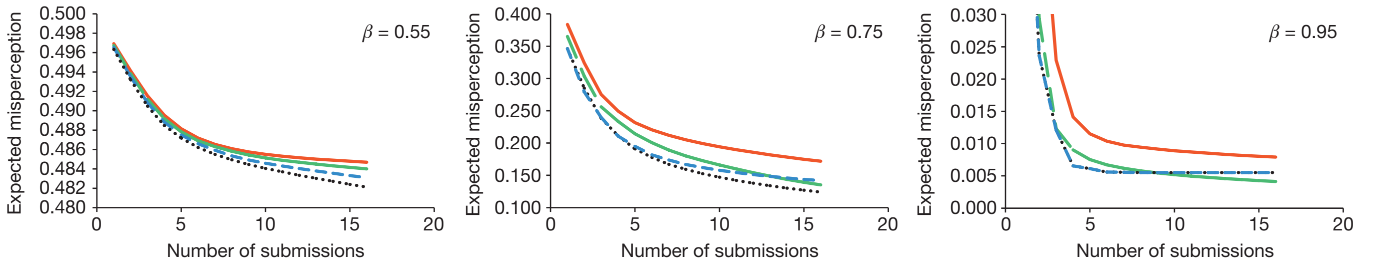


Figure 2 | Expected misperception in a generalized version of the M1 model. We show the expected misperception for three values of the probability that the initial opinion is correct (β): (1) 0.55 (left), (2) 0.75 (middle), and (3) 0.95 (right), reflecting high, intermediate, and low uncertainty. Results are shown for differing degrees to which the reviewer's subjective assessment determines their recommendation (ν): (1) 0.75 (red, solid line), (2) 1.00 (green, long dashed line), (3) 1.25 (blue, short dashed line), and (4) 1.50 (black, dotted line). In the original M1 model $\nu = 1$, while lower values reflect a more objective reviewer, and higher values a more subjective reviewer. Excessively subjective reviews are not effective in restraining incorrect herding (this is not yet visible for $\beta = 0.55$, but would become apparent with more submissions).

the accumulated information, rather than their original opinion, as publication history lengthens. In other words, in this case even the reviewers' recommendations are subject to herding. It appears that a moderate degree of subjectivity (as depicted in M1) is near-optimal.

Two empirical examples show that herding occurs in the scientific literature. First, belief in a specific scientific claim can be (and is) distorted through preferential citations of studies which support a particular point of view rather than those which do not[17]. This phenomenon can be attributed to herding caused by preferential citations, potentially creating a spurious and unfounded sense of authority for specific claims. Second, using a meta-analytic review of a recent literature[18], we compared claims made in the abstracts of the contributing studies with support for those claims by the data reported therein. Meta-analysis imposes a standard analysis to maximize comparability, and thereby minimizes the extent to which the presentation of results can be influenced by flexible analytical options[19]. These results (Fig. 3) show a mismatch between the claims made in the abstracts, and the strength of evidence for those claims based on a neutral analysis of the data, consistent with the occurrence of herding.

We next consider whether scientists can decide on their conclusion before conducting an experiment. We suggest that herding leads to one outcome being preferable over another, and that flexible analysis and selective reporting allows data that do not conform to either be transformed[19] or relegated to the file drawer[20]. Mendel famously appears to have dropped observations from his data so that his results conformed to his expectations[21], but because his theory was ultimately proved correct this is now generally overlooked. There is in fact clear evidence that the reporting and interpretation of findings is often inconsistent with the actual results[22], and this appears to be particularly pronounced in abstracts of research articles (often the only part that is read)[23].

Scientists may be motivated by a number of factors, such as the desire to be the first to advocate an idea, and the natural tendency to side with others of a similar opinion. Herding is therefore expected when agents care only about being published and recognize some topics as 'hot' (and therefore publishable). If scientists are motivated in this way in our model, then in an equilibrium of the game they can simply follow the first author's claim to maximize the probability of being published (see Supplementary Information). However, our results indicate that we can expect herding, including convergence on false conclusions, even when scientists—both as authors and reviewers—are rational and motivated by the pursuit of truth. The emergence of fads and fashions in the scientific literature (that is, hot topics)[1] is therefore unsurprising.

The first herding model in economics modelled individuals' investment choices[24]. Herding may have positive consequences, by driving rapid convergence on a correct decision. Rational individuals process all the information available to them before making decisions, and herding therefore arises from natural motives—a rational individual in pursuit of truth can and should be influenced by what others think. That humans are influenced in this way has been shown by experiments in social psychology[25]. It is rational because humans are aware of their own fallibility, and so their opinions may be strengthened or weakened by the views of others. In other words, being aware of the wisdom of the crowd, humans are (rationally) influenced by the crowd; in order to update our beliefs in the light of new evidence, we should be guided by Bayes' theorem. However, herding may also have negative consequences, by driving convergence on an incorrect decision. This is particularly problematic if an outsider to the process is unaware that it is taking place, as it gives a spurious sense of certainty to the observed convergence.

Free, open and global access to research reports has been proposed as an alternative to peer review (http://am.ascb.org/dora/), but, as we have shown, peer review can reveal more information relative to free and complete sequential publication. Reviewer recommendations, and resulting editor decisions, contain information, and thus prevent herding from completely blocking new information flow. However, this depends on specific parameters such as the popularity of the subject (for example, how many people are writing about this issue, or how long it is discussed) and how strongly scientists feel about their initial dispositions (that is, the level of β). In particular, if reviewers (and editors) are explicitly encouraged to be as objective as possible they will not be guided by Bayes' theorem when making their recommendations—it is only when reviewers are allowed a degree of subjectivity that this is done. Our results indicate that peer review performs best when the reviewers exercise their subjectivity at an intermediate level; higher levels enhance the risk of complete herding in reviewer decisions, whereas lower levels curb the information flow from reviewer decisions.

The peer-review process is therefore in principle self-correcting over a sufficiently extended period (although distortions may occur in the shorter term), in that de-herding can also occur. In reality, de-herding will not always occur, because publication histories within a topic may not

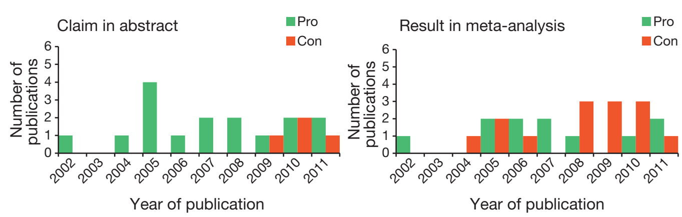


Figure 3 | Empirical evidence of discrepancy between claims and results. We show claims made in the abstracts of studies, and the results of those studies derived from a standardized analysis. Abstracts were coded as pro or con depending on whether an association was claimed, based on the judgement of an independent rater. Results were coded as pro or con depending on whether the overall effect size for the full sample in the study was statistically significant at $P < 0.05$. Five abstracts could not be coded as either pro or con. The proportion of pro versus con classifications differed for claims (80% pro) and results (44% pro), suggesting herding around the first published claim (McNemar test: $P = 0.016$, two-tailed test). Treating abstracts that could not be coded as pro or con did not alter these results substantially (84% versus 64% pro).

persist for sufficiently long. Science may therefore not be as self-correcting as is commonly assumed[8], and peer-review models which encourage objectivity over subjectivity may reduce the ability of science to self-correct. Although herding among agents is well understood in cases where the incentives directly reward acting in accord with the crowd (for example, financial markets), it is instructive to see that it can occur when agents (that is, scientists) are motivated by the pursuit of truth, and when gatekeepers (that is, reviewers and editors) exist with the same motivation. In such cases, it is important that individuals put weight on their private signals, in order to be able to escape from herding. Behavioural economic experiments indicate that prediction markets, which aggregate private signals across market participants, might provide information advantages[26]. Knowledge in scientific research is often highly diffuse, across individuals and groups[26], and publishing and peer-review models should attempt to capture this. We have discussed the importance of allowing reviewers to express subjective opinions in their recommendations, but other approaches, such as the use of post-publication peer review, may achieve the same end.

METHODS SUMMARY

Model. A number of scientists, indexed as $i = 1, 2, \cdots$, deliberate over two opposing hypotheses perceived *ex ante* to be equally likely to be true. Initially each scientist i receives an independent private signal regarding the true hypothesis, which is correct with probability $\beta \in \left(\frac{1}{2}, 1\right)$. Sequentially, scientist i submits a manuscript defending one of the two hypotheses, termed its theme, which is reviewed by the next scientist $i+1$ who decides whether to accept or reject the manuscript. This decision, and the theme if accepted, becomes common knowledge. Each scientist submits a manuscript defending a theme that is more likely to be the true hypothesis according to their posterior belief, formed by Bayes' rule based on all the information available at that time. We consider three models of reviewer decision. In M1, the reviewer accepts a manuscript with a probability proportional to the likelihood of its theme being true according to their posterior belief. In M2, they accept it irrespective of its theme with the *ex ante* probability they would accept a manuscript after the same publication history in M1. In M3, they simply accept it.
Concepts. A scientist is herding if their posterior belief attaches a probability greater than 0.5 to a particular hypothesis regardless of their own signal when they submit. Their probability of herding is the *ex ante* probability that they will be herding. The misperception after a publication history is the expected probability attached to the hypothesis, which is in reality incorrect, by outside observers who form their posterior beliefs on true hypothesis by Bayes' rule based on the history. The expected misperception after n submissions is the probability-weighted sum of misperceptions over all possible histories that may occur with n submissions.
Analysis. We wrote a computer program to recursively calculate numerical values of algebraic formulae for various concepts reported, and algebraically derived asymptotic properties for large numbers of submissions.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 August; accepted 16 October 2013.
Published online 4 December 2013.

1. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).
2. Ioannidis, J. P. A. Scientific inbreeding and same-team replication: Type D personality as an example. *J. Psychosom. Res.* **73,** 408–410 (2012).
3. Ioannidis, J. P. A. Contradicted and initially stronger effects in highly cited clinical research. *J. Am. Med. Assoc.* **294,** 218–228 (2005).
4. Ioannidis, J. P. & Trikalinos, T. A. Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *J. Clin. Epidemiol.* **58,** 543–549 (2005).
5. Davey Smith, G. in *Biopsychosocial Medicine: An Integrated Approach to Understanding Illness* (ed. White, P.) 77–102 (Oxford Univ. Press, 2005).
6. Tatsioni, A., Bonitsis, N. G. & Ioannidis, J. P. A. Persistence of contradicted claims in the literature. *J. Am. Med. Assoc.* **298,** 2517–2526 (2007).
7. Freese, J. in *Intergenerational Caregiving* (eds Crouter A. C., Booth A., Bianchi S. M. & Seltzer J. A.) 145–177 (Urban Institute Press, 2008).
8. Ioannidis, J. P. A. Why science is not necessarily self-correcting. *Perspect. Psychol. Sci.* **7,** 645–654 (2012).
9. Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* **54,** 30–34 (1959).
10. Barnes, B., Bloor, D. & Henry, J. *Scientific Knowledge: A Sociological Analysis*. (Univ. Chicago Press, 1996).
11. Kuhn, T. S. *The Structure of Scientific Revolutions*. (Univ. Chicago Press, 1962).
12. Yong, E. & Simonsohn, U. The data detective. *Nature* **487,** 18–19 (2012).
13. Martinson, B. C., Anderson, M. S. & de Vries, R. Scientists behaving badly. *Nature* **435,** 737–738 (2005).
14. Pfeiffer, T. & Hoffmann, R. Large-scale assessment of the effect of popularity on the reliability of research. *PLoS ONE* **4,** e5996 (2009).
15. Brembs, B., Button, K. & Munafò, M. R. Deep impact: unintended consequences of journal rank. *Front. Hum. Neurosci.* **7,** 291 (2013).
16. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14,** 365–376 (2013).
17. Greenberg, S. A. How citation distortions create unfounded authority: analysis of a citation network. *Br. Med. J.* **339,** b2680 (2009).
18. Murphy, S. E. *et al.* The effect of the serotonin transporter polymorphism (5-HTTLPR) on amygdala function: a meta-analysis. *Mol. Psychiatry* **18,** 512–520 (2013).
19. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22,** 1359–1366 (2011).
20. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol. Bull.* **86,** 638–641 (1979).
21. Edwards, A. W. F. More on the too-good-to-be-true paradox and Gregor Mendel. *J. Hered.* **77,** 138 (1986).
22. Boutron, I., Dutton, S., Ravaud, P. & Altman, D. G. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *J. Am. Med. Assoc.* **303,** 2058–2064 (2010).
23. Gøtzsche, P. C. Believability of relative risks and odds ratios in abstracts: cross sectional study. *BMJ* **333,** 231–234 (2006).
24. Banerjee, A. V. A simple model of herd behavior. *Q. J. Econ.* **107,** 797–817 (1992).
25. Asch, S. E. Studies of independence and conformity. *Psychol. Monogr.* **70,** 1–70 (1956).
26. Almenberg, J., Kittlitz, K. & Pfeiffer, T. An experiment on prediction markets in science. *PLoS ONE* **4,** e8500 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was supported by an Economics and Social Research Council UK PhD studentship to M.W.P. M.R.M. is a member of the UK Centre for Tobacco and Alcohol Studies, a UKCRC Public Health Research Centre of Excellence. Funding from the British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. The authors are grateful to S. Murphy for her assistance in coding the meta-analysis study abstracts, and to A. Bird and G. Huxley for their comments on earlier drafts of this manuscript.

Author Contributions All authors contributed equally to the design and analysis of the models and the writing of the manuscript. The project was conceived by I.-U.P. and M.R.M., and the computer program was written by M.W.P.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.R.M. (marcus.munafo@bristol.ac.uk).

METHODS

Model of the peer review process. We analyse a model in which *n* + *1 ex ante* identical scientists deliberate over two opposing hypotheses, labelled A and B. It is known that only one of these hypotheses is correct, and that *ex ante* both are equally likely to be correct. Denoting the correct hypothesis by τ, this is expressed as $P(\tau=A)=P(\tau=B)=\frac{1}{2}$. Before the game starts, each scientist i receives a private signal, $s_i \in \{A,B\}$, regarding which is the true hypothesis. These signals are independent random variables that assume a value equal to the correct hypothesis with probability β. The signals are informative but not perfect, that is, $\beta \in \left(\frac{1}{2},1\right)$. Lower values of β can be interpreted as reflecting a more controversial nature of the issue under question, when the signals tend to be less accurate.

Sequentially, and motivated to publish what is true, different scientists submit a manuscript, each defending a particular hypothesis. The 'theme' of scientist i's manuscript, $t_i \in \{A,B\}$, denotes the hypothesis that is defended. We postulate that, upon receiving a manuscript, the editor elicits peer review from a scientist whose stance on the topic is unknown to the editor, which eliminates the editor's influence on the editorial decision through reviewer selection. This is done to focus our analysis on reviewer behaviour, and means that in our model each manuscript is assigned to a scientist who has neither submitted their own manuscript nor acted as a reviewer at that point (because otherwise the editor would have inference on their stance from the theme of their submission or their previous decision as a reviewer). The editor follows the reviewer's recommendation in deciding whether to accept or reject the manuscript. If it is accepted, its theme becomes common knowledge; if it is rejected, the theme is not disclosed, but the rejection becomes common knowledge. Then, a new submission is made by a scientist who has not submitted before. In particular, our analysis is focused on the case that the next scientist who submits a manuscript is the one who reviewed the previous manuscript.

Thus, labelling the scientist who writes the i-th submission as i, each scientist $i \in \{1,2,\ldots,n\}$ sequentially submits a manuscript advocating a theme $t_i \in \{A,B\}$, which is reviewed by the next scientist $j=i+1$, who subsequently writes and submits their own manuscript. Scientist *n* + *1*, who also receives a signal s_{n+1}, only reviews. Scientists observe the history of publication outcomes as they arise. Let $h^i \in \{A,B,\varnothing\}^i$ denote a history of the first i publication outcomes, where each published manuscript is recorded by its theme, A or B, and each unpublished manuscript by Ø. Then, there are three items of information available to each scientist j when they make decisions: (1) their own private signal $s_j \in \{A,B\}$; (2) a manuscript to be reviewed with a theme $t_{j-1} \in \{A,B\}$ if $j>1$; and (3) a history $h^{j-2} \in \{A,B,\varnothing\}^{j-2}$ if $j>2$. The two decisions to make are whether or not to recommend acceptance of a manuscript that they are reviewing, and the theme of the manuscript they subsequently submit.

We made a few modelling choices that simplify real practices, namely that: (1) only one reviewer is consulted for each submission; (2) the current reviewer is the next author; (3) rejections become common knowledge; and, (4) authors conform to the rationality assumption that they are Bayesian updaters. Choices 1 and 2 maximize the number of submissions that can be reviewed by a given number of scientists, subject to the editor not soliciting a review from someone with a known stance. Choice 3 spares scientists from having to make probabilistic inferences as to what other submissions might have been made but rejected, which would have been necessary to determine the optimal choices when they act. These features enable us to examine the largest possible number of submissions with the available computing power, and thus allow us to generate more meaningful outputs without changing the essential processes operating. We believe that our main message will remain valid when these assumptions are relaxed (see Supplementary Information for a further discussion of choice 3). However, the complexity of the computer program needed to analyse such cases, and the corresponding computing power required, will increase exponentially. Choice 4 assumes authors use all of the information available to them, in accordance with Bayes' theorem[27], to determine the relative likelihood (called a posterior belief) that each of the two alternative hypotheses is correct. Then, being motivated to publish what is true, each scientist will submit a manuscript advocating the hypothesis that is more likely to be correct according to their posterior belief, augmented by a standard tie-breaking rule of following their own signal when both are equally likely[24]. This is one of the rationality assumptions that economists place on humans.

Models of reviewer behaviour. In the first model, M1, scientist $j=i+1$ recommends acceptance of scientist i's manuscript with the same probability, denoted by $P(\tau=t_i|\beta,h^{j-2},t_i,s_j)$, that they infer the theme of the manuscript to be the correct hypothesis, by Bayes' rule based on all the information available to them at that point. Therefore, reviewers as well as authors act guided by Bayesian inference in this model. The acceptance probabilities are endogenous and evolve differently depending on how the publication history unfolds.

In the second model, M2, the acceptance decision is completely independent of the reviewer's subjective assessment of the theme of the manuscript, and rather is based on other, largely objective characteristics of the manuscript, such as the quality of the research methodology. Presuming that these traits are statistically independent of the manuscript's conclusion, the acceptance probabilities in M2 are independent of both the theme of the manuscript and the assigned reviewer (insofar as the only feature that distinguishes reviewers is their assessment of which hypothesis is correct). Thus, the acceptance probabilities can be thought of as the likelihood that the methodological quality of the manuscript is sufficient to warrant publication, and not a reflection of whether or not the reviewer agrees with the conclusions. However, our model does not specify what those probabilities should be. To aid comparison between the models, we considered two cases. In one, scientist j, irrespective of their own signal, recommends acceptance of i's manuscript with a probability equal to the *ex ante* probability that they would recommend acceptance of i's manuscript in M1 after the same history (this results in the same expected number of publications in both M1 and M2). In the other, the acceptance probability remains the same throughout, at the initial expected acceptance probability of the M1 model, which is β. To verify this, note that scientist 2 would recommend acceptance of scientist 1's manuscript with probability $\frac{\beta^2}{\beta^2+(1-\beta)^2}$ when s_2 agrees with t_1 (which happens with probability $\beta^2+(1-\beta)^2=1-2\beta+2\beta^2$) but with probability 0.5 otherwise. Hence, the expected probability of acceptance is $\beta^2+\frac{1}{2}\left(2\beta-2\beta^2\right)=\beta$. As the results are similar in the two cases of M2, here we report only on the former.

In the third (benchmark) model, M3, all manuscripts are published without any filtering through peer review. This model is identical to M2 but with the acceptance probability equal to 1 throughout the process. This is a simple model of herd behaviour[24,28] that has become standard in economics when modelling self-motivated, rational individuals who sequentially take actions. A consequence of this model is that each scientist will have access to all previous submissions when forming their decision (because everything is published in this model). Note that this differs from a full information case (that is, where every scientist has access to all private signals, as well as public actions).

In the generalized M1 models, scientist j recommends acceptance with probability $\min\left\{1,\frac{1}{2}+v\cdot\left(P\left(\tau=t_i|\beta,h^{j-2},t_i,s_j\right)-\frac{1}{2}\right)\right\}$ if $P\left(\tau=t_i|\beta,h^{j-2},t_i,s_j\right)\geq\frac{1}{2}$, and with probability $\max\left\{0,\frac{1}{2}+v\cdot\left(P\left(\tau=t_i|\beta,h^{j-2},t_i,s_j\right)-\frac{1}{2}\right)\right\}$ if $P(\tau=t_i|\beta,h^{j-2},t_i,s_j)<\frac{1}{2}$, where $v>0$. The case $v=1$ corresponds to the original M1 model, with higher values of v indicating that the recommendation is more heavily influenced by the reviewer's subjective assessment on the advocated theme, and lower v meaning that it is less so.

Definitions and algebraic formulae. The misperception is defined from the perspective of outsiders who observe the publication history. Using all the information available to them from the observed history, $h^n \in \{A,B,\varnothing\}^n$, outside observers will form via Bayes' rule a posterior belief that attaches probability $P(\tau|h^n)=\frac{P(h^n|\tau)}{P(h^n|A)+P(h^n|B)}$ to hypothesis τ being true for $\tau \in \{A,B\}$, where $P(h^n|\tau)$ is the probability that the history h^n realizes under hypothesis $\tau \in \{A,B\}$. We define the misperception, after history h^n, as the expected posterior probability attached to the hypothesis which is in reality incorrect: since $P(\tau=A)=P(\tau=B)=\frac{1}{2}$, it is:

$$\frac{\frac{1}{2}\sum_{\tau=A,B}[1-P(\tau|h^n)]\cdot P(h^n|\tau)}{\frac{1}{2}\sum_{\tau=A,B}P(h^n|\tau)} \tag{1}$$

The expected misperception after n submissions is defined as a probability-weighted sum of misperceptions over all possible histories of length n that may occur:

$$E[\text{misperception}]=\frac{1}{2}\sum_{\tau=A,B}\sum_{h^n\in\{A,B,\varnothing\}^n}[1-P(\tau|h^n)]P(h^n|\tau) \tag{2}$$

Note that these calculations are done for an underlying value of β.

Focusing on h^1 (for which we need two scientists), there are three possible histories, namely $h^1 \in \{A,B,\varnothing\}$. Equation (2), above, which gives us the expected misperception, will have 6 terms when $n=1$, because each of the three histories can occur from either hypothesis $\tau \in \{A,B\}$. Note that $P(\tau|h^1)$ is symmetric in the sense that its value remains the same when A and B (as values of τ and elements of h^1) are permuted. A consequence of this symmetry is that we only need to consider the case when one hypothesis (for example, A) is correct, and the sum of 6 terms

will be equal to twice of the sum of the three items relevant for $\tau = A$. For $n = 2$ because there are $3^2 = 9$ possible histories, there will be 9 terms to calculate (after taking into account the symmetry). Similarly, the expected misperception after n submissions can be obtained by calculating 3^n terms:

$$E[\text{misperception}] = \sum_{h^n \in \{A,B,\varnothing\}^n} [1 - P(A|h^n)]P(h^n|A) \quad (3)$$

Herding is defined for scientists who are submitting papers. A scientist, say j, is said to be herding if they would choose the same theme to advocate regardless of their private signal as their posterior belief would attach a probability more than one half to a particular hypothesis regardless of their own signal, that is, if:

$$\text{For some } \tau \in \{A,B\},\ \min\{P(\tau|\beta,h^{j-2},t_{j-1},s_j = A), P(\tau|\beta,h^{j-2},t_{j-1},s_j = B)\} > \frac{1}{2} \quad (4)$$

The probability of herding, for a scientist j, can easily be calculated by the following probability-weighted sum:

$$\text{Probability of herding} = \sum_{\{\forall h^{j-2},t_{j-1}\}} 1_H(h^{j-2},t_{j-1}) \cdot P(h^{j-2},t_{j-1}) \quad (5)$$

where $P(h^{j-2}, t_{j-1})$is the probability that (h^{j-2}, t_{j-1}) realizes from either hypothesis and 1_H is the indicator function that assumes a value of 1 if (4) holds, and 0 otherwise.

When herding occurs, some histories and information profiles will occur with probability zero. This means that there will generally be a number of terms in (3) and (5) that will never occur, so the calculations required will generally be over a smaller number of terms than the theoretical upper bound. Nevertheless, the large number of terms that result from even a moderate n are impossible to simplify to obtain a closed-form algebraic expression for either the expected misperception or the probability of herding. We therefore wrote a computer program to numerically calculate the algebraic expressions within available computing power.

Computer program. The program (code provided in the Supplementary Information) worked by building and evaluating the algebraic formulae to obtain results that are accurate up to the level of precision the computer used in its calculations (52 dp), as explained through a number of key steps described below for various values of β. The information a reviewer j has, $(h^{j-2},t_{j-1},s_j) \in \{A, B, \varnothing\}^{j-2} \times \{A,B\}^2$, is referred to as their 'information profile'.

Step 1: For each of the two possible private signals of scientist 1, $s_1 \in \{A,B\}$, a probability is set for the occurrence of that signal conditional on each of the two hypothesis $\tau \in \{A,B\}$: $P(s_1|\tau) = \beta$ if $s_1 = \tau$ and $P(s_1|\tau) = 1 - \beta$ otherwise. Thus, the posterior on the true hypothesis is calculated as: $P(\tau|s_1) = \dfrac{P(s_1|\tau)}{P(s_1|A) + P(s_1|B)}$.

Step 2: For each signal s_1 a submission decision of scientist 1 is prescribed. As $P(\tau = s_1|s_1) > 0.5$, for scientist 1 the theme of their submitted paper (t_1) will be identical to their signal (s_1). This determines the probability of $t_1 \in \{A,B\}$ conditional on $\tau \in \{A,B\}$.

Step 3: For each possible information profile $(t_1,s_2) \in \{A,B\}^2$ of scientist 2, the probability of acceptance (of scientist 1's submission with theme t_1) is determined in accordance with the adopted model. For M1 (and hence, M2), this involves calculating scientist 2's posterior beliefs as $P(\tau|t_1,s_2) = \dfrac{P(t_1,s_2|\tau)}{P(t_1,s_2|A) + P(t_1,s_2|B)}$ where $P(t_1,s_2|\tau) = P(t_1|\tau)P(s_2|\tau)$.

Step 4: If scientist 1's manuscript is rejected, a history $h^1 = \varnothing$ ensues. If accepted, a history $h^1 = t_1$ ensues. For each possible history h^1, the conditional probability $P(h^1|\tau)$ is obtained by aggregating the probabilities that it arises from different signal profiles (s_1, s_2) conditional on τ. The misperception is calculated for each history according to the formula (1), and then the expected misperception according to the formula (3).

Step 5: The submission decision of scientist 2, t_2, is equal to τ such that $P(\tau|t_1,s_2) > 0.5$ if such a τ exists; otherwise, that is, if $P(A|t_1,s_2) = P(B|t_1,s_2) = 0.5$, then $t_2 = s_2$. This determines the conditional probability $P(h^1,t_2|\tau)$. Herding (and other results) is calculated according to the relevant formulae given.

Step 6: Steps 3–5 are repeated for $j \in \{3, \ldots n+1\}$ for every possible information profile (h^{j-2},t_{j-1},s_j) of scientist j with the following modifications: scientist j's posterior beliefs are $P(\tau|h^{j-2},t_{j-1},s_j) = \dfrac{P(h^{j-2},t_{j-1},s_j|\tau)}{P(h^{j-2},t_{j-1},s_j|A) + P(h^{j-2},t_{j-1},s_j|B)}$ where $P(h^{j-2},t_{j-1},s_j|\tau) = P(h^{j-2},t_{j-1}|\tau)P(s_j|\tau)$ in step 3; $h^{j-1} \in h^{j-2} \times \{A,B,\varnothing\}$ replaces h^1 and $P(h^{j-1}|\tau)$ is obtained by combining $P(h^{j-2},t_{j-1},s_j|\tau)$ and scientist j's acceptance probability given their information profile (h^{j-2},t_{j-1},s_j) in step 4; and $P(\tau|h^{j-2},t_{j-1},s_j)$ and $P(h^{j-1},t_j|\tau)$ replace $P(\tau|t_1,s_2)$ and $P(h^1,t_2|\tau)$, respectively, in step 5.

Analytical results on asymptotic properties. Analytic comparison of different models is obtained asymptotically as the numbers of scientists tends to infinity. Consider M1. Let $H^n = \{A,B,\varnothing\}^n$ denote the set of all possible histories of length n, and $h^n \in H^n$ denote a history in H^n. Then, $F_n = \{\varnothing\} \cup H^1 \cup \cdots \cup H^n$ for $n = 1,2,\cdots$, constitute an infinite sequence of σ-fields on H^∞.

For each h^n, let $P(h^n)$ be the *ex ante* probability that h^n will realize from either $\tau = A,B$. Let $X_n(h^n) = \dfrac{P(h^n|A)}{P(h^n|A) + P(h^n|B)}$ denote the Bayes-updated posterior belief that $\tau = A$ after h^n. Then, X_n is a random variable defined on (H^∞,F_n,P), and $\{(X_n,F_n)\}_{n=1,2,\cdots}$ constitutes a martingale. Let $Q(h^n) = P(h^n|A)$. Then, with X_n defined on (H^∞,F_n,Q), the sequence $\{(X_n,F_n)\}_{n=1,2,\cdots}$ constitutes a submartingale. By the Martingale Convergence Theorem[29], $E(X_n) \to E(X)$ almost surely where X is a random variable such that $X_n \to X$ with probability 1 and $E(\cdot)$ is taken relative to Q.

Consider a history h^n with the corresponding posterior $X_n = x < 1$. Then, there are three possible continuation histories of length $n + 1$: h^n followed by A, B, or Ø. As the manuscript of scientist $n + 1$ is accepted with a probability that is strictly between 0 and 1, (i) at least two of the three possible continuation histories realize with a strictly positive probability. Furthermore, (ii) the posteriors after these continuation histories differ, (iii) they depend on x but not on n, (iv) the distribution over these posteriors conditional on τ = A first-order stochastically dominates that conditional on $\tau = B$. Hence, $E(X_{n+1}|X_n = x) - x$ is a strictly positive constant that depends on x but not on n, and consequently, $E(X) < 1$ is not viable. As $E(X) \leq 1$, therefore, we conclude that $E(X) = 1$, that is, the posterior converges to true state with probability 1 when $\tau = A$. As a symmetric argument applies to the case that $Q(h^n) = P(h^n|B)$, that is, when $\tau = B$, the misperception converges to 0 as $n \to \infty$ under M1.

Next, consider the generalized M1 model with $\nu > 0$. As long as $0 < \nu < 1$, it is straightforward to verify that the deductions (i)–(iv) hold and, consequently, the same argument as above leads to the same conclusion that the misperception converges to 0 as $n \to \infty$. If $\nu > 1$, on the other hand, any manuscript on theme τ will be accepted with certainty once the posterior belief for the theme being true exceeds a certain threshold level which is strictly below 1. In addition, the scientists will submit on the popular theme regardless their own signal if the posterior for that theme exceeds a (different) threshold. Therefore, if the posterior belief for $\tau = A$ gets sufficiently close to 1 or 0, both the author's theme selection and the reviewer's decision are uniquely determined by the prevailing posterior independently of the scientist's own signal. Once this stage is reached, then the continuation history is uniquely determined (irrespective of whether $\tau = A$ or B) unlike (i) above and, consequently, publication outcomes reveal no further information and the posterior remains at the same level forever. Therefore, the expected misperception never converges to 0 and remains fixed at some positive level within finite time with probability 1.

For M2 and M3, by the same token the expected misperception never converges to 0 and gets stuck at some positive level once the posterior belief reaches a level such that the author's theme selection is dictated by herding independently of their own signal.

27. Bayes, T. & Price, R. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions (1683–1775)* **53,** 370–418 (1763).
28. Bikhchandani, S., Hirshleifer, D. & Welch, I. A theory of fads, fashion, custom, and cultural change in informational cascades. *J. Polit. Econ.* **100,** 992–1026 (1992).
29. Billingsley, P. *Probability and Measure* 3rd edn (John Wiley & Sons, 1995).

doi:10.1038/nature12828

Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico

The SIGMA Type 2 Diabetes Consortium*

Performing genetic studies in multiple human populations can identify disease risk alleles that are common in one population but rare in others[1], with the potential to illuminate pathophysiology, health disparities, and the population genetic origins of disease alleles. Here we analysed 9.2 million single nucleotide polymorphisms (SNPs) in each of 8,214 Mexicans and other Latin Americans: 3,848 with type 2 diabetes and 4,366 non-diabetic controls. In addition to replicating previous findings[2–4], we identified a novel locus associated with type 2 diabetes at genome-wide significance spanning the solute carriers *SLC16A11* and *SLC16A13* ($P = 3.9 \times 10^{-13}$; odds ratio (OR) = 1.29). The association was stronger in younger, leaner people with type 2 diabetes, and replicated in independent samples ($P = 1.1 \times 10^{-4}$; OR = 1.20). The risk haplotype carries four amino acid substitutions, all in SLC16A11; it is present at ~50% frequency in Native American samples and ~10% in east Asian, but is rare in European and African samples. Analysis of an archaic genome sequence indicated that the risk haplotype introgressed into modern humans via admixture with Neanderthals. The *SLC16A11* messenger RNA is expressed in liver, and V5-tagged SLC16A11 protein localizes to the endoplasmic reticulum. Expression of SLC16A11 in heterologous cells alters lipid metabolism, most notably causing an increase in intracellular triacylglycerol levels. Despite type 2 diabetes having been well studied by genome-wide association studies in other populations, analysis in Mexican and Latin American individuals identified *SLC16A11* as a novel candidate gene for type 2 diabetes with a possible role in triacylglycerol metabolism.

The Slim Initiative in Genomic Medicine for the Americas (SIGMA) Type 2 Diabetes Consortium set out to characterize the genetic basis of type 2 diabetes in Mexican and other Latin American populations, where the prevalence is roughly twice that of US non-Hispanic whites[5] (see also http://www.cdc.gov/diabetes/pubs/factsheet11.htm). This report considers 3,848 type 2 diabetes cases and 4,366 controls (Table 1) genotyped using the Illumina OMNI 2.5 array that were unrelated to other samples, and that fall on a cline of Native American and European ancestry[6] (Extended Data Fig. 1). Association analysis included 9.2 million variants that were imputed[7,8] from the 1000 Genomes Project Phase I release[9] based on 1.38 million SNPs directly genotyped at high quality with minor allele frequency (MAF) >1%.

The association of SNP genotype with type 2 diabetes was evaluated using LTSOFT[10], a method that increases power by jointly modelling case–control status with non-genetic risk factors. Our analysis used body mass index (BMI) and age to construct liability scores and also included adjustment for sex and ancestry via principal components[6]. The quantile–quantile (QQ) plot is well calibrated under the null ($\lambda_{GC} = 1.05$; Fig. 1a, red), indicating adequate control for confounders, with substantial excess signal at $P < 10^{-4}$.

We first examined SNPs previously reported to be associated to risk of type 2 diabetes. Two such variants reached genome-wide significance: *TCF7L2* (rs7903146; $P = 2.5 \times 10^{-17}$; OR = 1.41 (95% confidence interval 1.30–1.53)) and *KCNQ1* (rs2237897; $P = 4.9 \times 10^{-16}$; OR = 0.74 (0.69–0.80)) (Extended Data Figs 2, 3a), with effect sizes and frequencies consistent with previous studies[3,4,11]. At *KCNQ1*, we identified a signal[3] of association that shows limited linkage disequilibrium both to rs2237897 ($r^2 = 0.056$) and to rs231362 ($r^2 = 0.028$) (previously seen in Europeans[11]), suggesting a third allele at this locus (rs139647931; after conditioning, $P = 5.3 \times 10^{-8}$; OR = 0.78 (0.70–0.86); Extended Data Fig. 3b and Supplementary Note).

More generally, of SNPs previously associated with type 2 diabetes at genome-wide significance, 56 of 68 are directionally consistent with the initial report ($P = 3.1 \times 10^{-8}$; Supplementary Table 1). Nonetheless, a QQ plot excluding all SNPs within 1 megabase (Mb) of the 68 type 2 diabetes associations remains strikingly non-null (Fig. 1a, blue).

This excess signal of association is entirely attributable to two regions of the genome: chromosome 11p15.5 and 17p13.1 (Fig. 1a, black). The genome-wide significant association at 11p15.5 spans insulin, *IGF2* and other genes (Extended Data Fig. 3a): the SNP with the strongest association lies in the 3′ untranslated region (UTR) of *IGF2* and the non-coding *INS–IGF2* transcript (rs11564732, $P = 2.6 \times 10^{-8}$; OR = 0.77 (0.70–0.84); Supplementary Table 2). The associated SNPs are ~700 kilobases (kb) from the genome-wide significant signal in *KCNQ1* (above), and analysis conditional on the two significant *KCNQ1* SNPs reduced the *INS–IGF2* association signal to just below genome-wide significance ($P = 7.5 \times 10^{-7}$, Extended Data Fig. 3c). Conditioning on the two *KCNQ1* SNPs and the *INS–IGF2* SNP reduces the signal to background (Extended Data Fig. 3d). Further analysis is needed to determine whether the *INS–IGF2* signal is reproducible and independent of that at *KCNQ1*.

Table 1 | Study cohorts comprising the SIGMA type 2 diabetes project data set

Study	Sample location	Study design		*n* (before quality control)	Per cent male	Age (years)	Age-of-onset (years)	BMI (kg m^{-2})	Fasting plasma glucose (mmol l^{-1})
UNAM/INCMNSZ Diabetes Study (UIDS)	Mexico City, Mexico	Prospective cohort	Controls	1,138 (1,195)	41.1	55.3 ± 9.4	–	28.1 ± 4.0	4.8 ± 0.5
			T2D cases	815 (872)	40.9	56.2 ± 12.3	44.2 ± 11.3	28.4 ± 4.5	–
Diabetes in Mexico Study (DMS)	Mexico City, Mexico	Prospective cohort	Controls	472 (505)	25.8	52.5 ± 7.7	–	28.0 ± 4.4	5.0 ± 0.4
			T2D cases	690 (762)	33.0	55.8 ± 11.1	47.8 ± 10.6	29.0 ± 5.4	–
Mexico City Diabetes Study (MCDS)	Mexico City, Mexico	Prospective cohort	Controls	613 (790)	39.3	62.5 ± 7.7	–	29.4 ± 4.8	5.0 ± 0.5
			T2D cases	287 (358)	41.1	64.2 ± 7.5	55.1 ± 9.7	29.9 ± 5.4	–
Multiethnic Cohort (MEC)	Los Angeles, California, USA	Case–control	Controls	2,143 (2,464)	48.3	59.3 ± 7.0	–	26.6 ± 3.9	N/A
			T2D cases	2,056 (2,279)	47.9	59.2 ± 6.9	N/A	30.0 ± 5.4	–

The table shows sample location, study design, numbers of cases and controls (including numbers before quality control checks), per cent male participants, age ± standard deviation (s.d.), age-of-onset in cases ± s.d., body mass index ± s.d., and fasting plasma glucose in controls ± s.d. N/A, not applicable; T2D, type 2 diabetes.

*Lists of participants and their affiliations appear at the end of the paper.

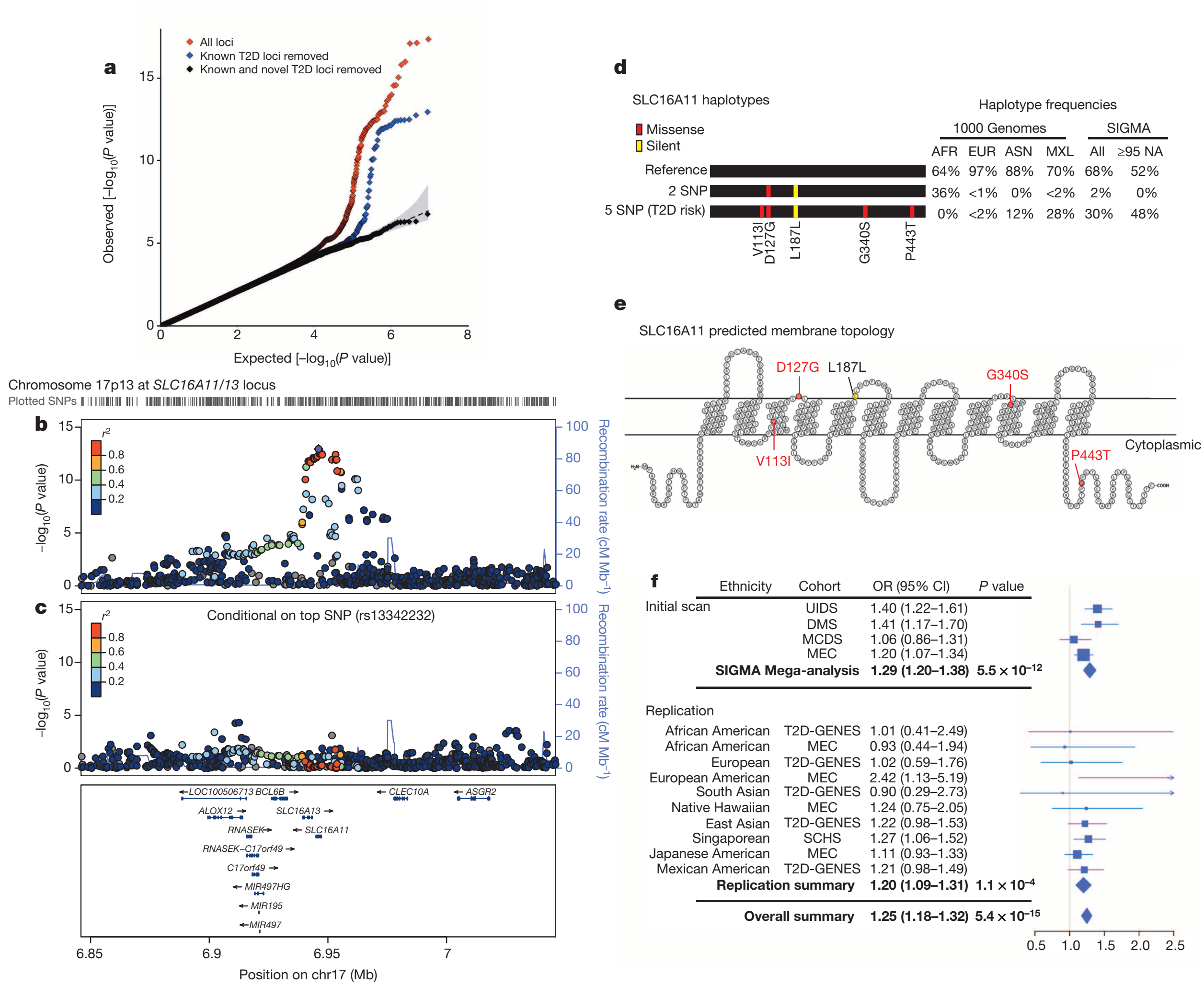


Figure 1 | Identification of a novel type 2 diabetes risk haplotype carrying 5 SNPs in *SLC16A11*. **a**, QQ plot of association statistics in genome-wide scan of $n = 8{,}214$ samples shows calibration under the null and enrichment in the tail for all SNPs (red), and after removing SNPs within 1 Mb of previously published type 2 diabetes associations (blue). Removal of sites within 1 Mb of 68 known loci and two novel loci results in a null distribution (black). Association with liability threshold quantitative traits tested via linear regression. T2D, type 2 diabetes. **b**, Regional plot of association at 17p13.1 that spans *SLC16A11* and *SLC16A13*. **c**, Analysis conditional on genotype at rs13342232 (the top associated variant) reduces signal to far below genome-wide significance across the surrounding region. Colour indicates r^2 to the most strongly associated site; recombination rate is shown, each based on the 1000 Genomes ASN population. **d**, Graphical depictions of SLC16A11 haplotypes constructed from the synonymous and four missense SNPs associated to type 2 diabetes, with haplotype frequencies derived from the 1000 Genomes Project and SIGMA samples. AFR, African ($n = 185$); ASN, east Asian ($n = 286$); EUR, European ($n = 379$); MXL, Mexican samples from Los Angeles ($n = 66$). Frequencies from SIGMA samples are calculated from genotypes and represent either the entire data set (All) or only samples estimated to have ≥95% Native American ancestry (≥95 NA, $n = 290$; Supplementary Methods). Haplotypes with population frequency <1% are not depicted. **e**, Predicted membrane topology of human SLC16A11 generated using TMHMM 2.0 and visualized with TeXtopo. Locations of SNPs carried by the type-2-diabetes-associated haplotype are indicated. **f**, Forest plot depicting odds ratio estimates at rs75493593 from the four SIGMA cohorts, the SIGMA pooled mega-analysis, the replication cohorts, replication-only meta-analysis based on inverse standard error weighting of effect sizes, and the overall meta-analysis (including all replication cohorts and the SIGMA mega-analysis). Accompanying table lists ethnicity, cohort names, estimated odds ratio (OR) and 95% confidence interval (95% CI). Replication cohorts are the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES), Multiethnic Cohort (MEC), and Singapore Chinese Health Study (SCHS). Further details including sample sizes are provided in Supplementary Table 8.

The strongest novel association is at 17p13.1 spanning *SLC16A11* and *SLC16A13* (Fig. 1b), both poorly characterized members of the monocarboxylic acid transporter family of solute carriers[12]. The strongest signal of association includes a silent mutation as well as four missense SNPs, all in *SLC16A11* (Fig. 1d, e). These five variants are (1) in strong linkage disequilibrium ($r^2 \geq 0.85$ in 1000 Genomes samples from the Americas) and co-segregate on a single haplotype; (2) common in samples of Latin American ancestry; and (3) show equivalent levels of association to type 2 diabetes ($P = 2.4 \times 10^{-12}$ to $P = 3.9 \times 10^{-13}$; OR = 1.29 (1.20–1.38); Supplementary Tables 3–5). Analysis conditional on any of these variants leaves no genome-wide significant signal (Fig. 1c and Extended Data Fig. 4). Computational prediction with SIFT[13] (which considers each site independently) labels one of the missense SNPs (rs13342692, D127G) as damaging and the other three 'tolerated' (Supplementary Table 6).

Individuals that carry the risk haplotype develop type 2 diabetes 2.1 years earlier ($P = 3.1 \times 10^{-4}$), and at 0.9 kg m^{-2} lower BMI ($P = 5.2 \times 10^{-4}$) than non-carriers (Extended Data Fig. 5). The odds ratio for the risk haplotype estimated using young cases (≤45 years) was higher than in older cases (OR = 1.48 versus 1.11; $P_{heterogeneity} = 1.7 \times 10^{-3}$). We tested the haplotype for association with related metabolic quantitative traits in the fasting state in a subset of SIGMA participants ($n = 1{,}505$–$3{,}855$). No associations surpass nominal significance ($P < 0.05$; Supplementary Table 7).

Given that large genome-wide association studies (GWAS) have been performed for type 2 diabetes in samples of European and Asian ancestry, it may seem surprising that associated variants at *SLC16A11/13* were not previously identified. Using data generated by the 1000 Genomes Project and the current study, we observed that the risk haplotype (hereafter referred to as '5 SNP' haplotype) is rare or absent in samples from Europe and Africa, has intermediate frequency (~10%) in samples from east Asia, and up to ~50% frequency in samples from the Americas (Fig. 1d and Extended Data Fig. 6a). A second haplotype carrying one of the four missense SNPs (D127G) and the synonymous variant (termed the '2 SNP' haplotype) is very common in samples from Africa but rare elsewhere, including in the Americas (Fig. 1d). The low frequency of the 5 SNP haplotype in Africa and Europe may explain why this association was not found in previous studies.

We attempted to replicate this association in ~22,000 samples from a variety of ancestry groups. A proxy for the 5 SNP haplotype of *SLC16A11* showed strong association with type 2 diabetes ($P_{replication} = 1.1 \times 10^{-4}$; $OR_{replication} = 1.20$ (1.09–1.31); $P_{combined} = 5.4 \times 10^{-15}$; $OR_{combined} = 1.25$ (1.18–1.32); Fig. 1f and Supplementary Table 8). The association was clearly observed in east Asian samples, a population that lacks admixture of Native American and European populations and shows little genetic substructure. This result argues against population stratification as an explanation for the finding in Latin American populations.

We estimated the difference in disease prevalence attributable to a risk factor with OR = 1.20 (1.09–1.31), 26% frequency in Mexican Americans (as in the SIGMA control samples) and 2% in European Americans. Approximately 20% (9.2–29%) of the difference in prevalence could be explained by such a risk factor (Supplementary Methods).

Two population genetic features of the 5 SNP haplotype struck us as discordant. The haplotype sequence is highly divergent, with an estimated time to most recent common ancestor (TMRCA) of 799,000 years to a European haplotype (Supplementary Table 9 and Supplementary Note). This long precedes the 'out of Africa' bottleneck. And yet, the haplotype is not observed in Africa and is rare throughout Europe (Fig. 1d).

This combination of age and geographical distribution could be consistent with admixture from Neanderthals into modern humans. Neither the published Neanderthal genome[14] nor the Denisova genome[15] contained the variants observed on the 5 SNP haplotype. However, an unpublished genome of a Neanderthal from Denisova Cave[16,17] is homozygous across 5 kb for the 5 SNP haplotype at *SLC16A11*, including all four missense SNPs. Over a span of 73 kb this Neanderthal sequence is nearly identical to that of individuals from the 1000 Genomes Project who are homozygous for the 5 SNP haplotype (Supplementary Note).

Two lines of evidence indicate that the 5 SNP haplotype entered modern humans through archaic admixture. First, the Neanderthal sequence is more closely related to the extended 73 kb 5 SNP haplotype than to random non-risk haplotypes (mean TMRCA = 250,000 years versus 677,000 years; Supplementary Tables 10 and 11 and Supplementary Note), forming a clade with the risk haplotype (Extended Data Fig. 6b) with a coalescence time that post-dates the range of estimated split times between modern humans and Neanderthals[15,18]. Second, the genetic length of the 73-kb haplotype is longer than would be expected if it had undergone recombination for ~9,000 generations since the split with Neanderthals ($P = 3.9 \times 10^{-5}$; Supplementary Note). These two features indicate that the 5 SNP haplotype is not only similar to the Neanderthal sequence, but was probably introduced into modern humans relatively recently through archaic admixture. We note that whereas this particular Neanderthal-derived haplotype is common in the Americas, Latin Americans have the same proportion of Neanderthal ancestry genome-wide as other Eurasian populations (~2%)[15].

With an absence of multiple independently segregating functional mutations in the same gene, we lack formal genetic proof that *SLC16A11* is the gene responsible for association to type 2 diabetes at 17p13.1. Nonetheless, as the associated haplotype encodes four missense SNPs in a single gene (Supplementary Table 12), we set out to begin characterizing the function of SLC16A11.

We examined the tissue distribution of *SLC16A11* mRNA expression using Nanostring and ~55,000 curated microarray samples. In both data sets, we observed *SLC16A11* expression in liver, salivary gland and thyroid (Extended Data Figs 7 and 8). We used immunofluorescence to determine the subcellular localization of V5-tagged SLC16A11 introduced into HeLa cells. SLC16A11–V5 co-localizes with the endoplasmic reticulum membrane protein calnexin, but shows minimal overlap with plasma membrane, Golgi apparatus and mitochondria (Fig. 2a). Distinct patterns were seen for other SLC16 family members, which are known to have diverse cellular functions[19]: SLC16A13–V5 localizes to the Golgi apparatus and SLC16A1–V5 appears at the plasma membrane[20] (Extended Data Fig. 9 and data not shown).

As SLC16 family members are solute carriers, we expressed SLC16A11 (or control proteins) in HeLa cells (which do not express SLC16A11 at

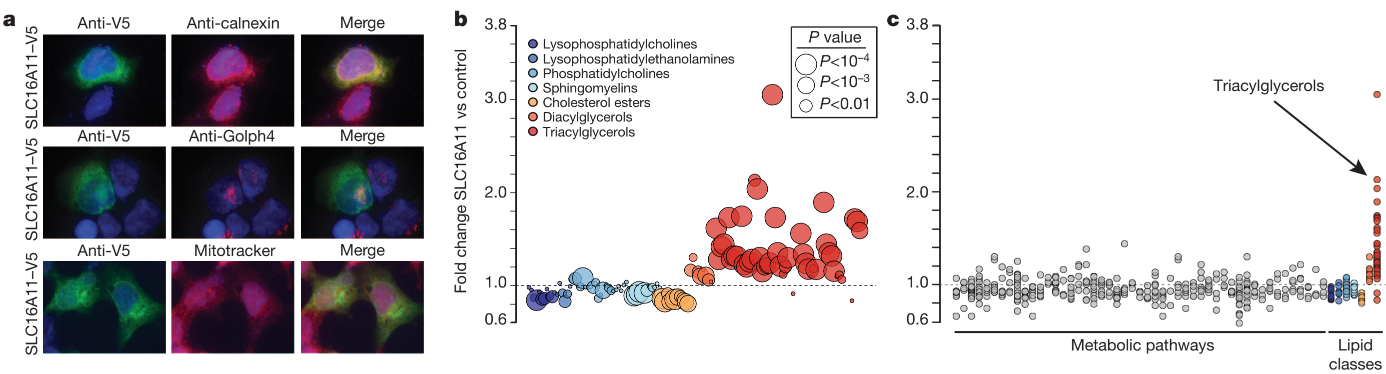


Figure 2 | SLC16A11 localizes to the endoplasmic reticulum and alters lipid metabolism in HeLa cells. a, Localization of SLC16A11 to the endoplasmic reticulum. HeLa cells expressing C terminus, V5-tagged SLC16A11 were immunostained for SLC16 expression (anti-V5) along with markers for the endoplasmic reticulum (anti-calnexin), cis-Golgi apparatus (anti-Golph4), or mitochondria (MitoTracker). Imaging of each protein was optimized for clarity of localization rather than comparison of expression level across proteins. Representative images from multiple independent transfections are shown. **b**, Changes in intracellular lipid metabolites after expression of SLC16A11–V5 in HeLa cells. The fold change in cells expressing SLC16A11 relative to cells expressing control proteins is plotted for individual lipid metabolites, with lipid classes indicated by point colour and *P* values (of the Wilcoxon rank-sum test) by point size. **c**, Fold change plotted for both polar and lipid metabolites, grouped according to metabolic pathway or class. Pathways shown include all KEGG pathways from the human reference set for which metabolites were measured as well as eight additional classes of metabolites covering carnitines and lipid subtypes. Each point within a pathway or class shows the fold change of a single metabolite within that pathway or class. Pathway names and statistical analyses are shown in Extended Data Fig. 10 and Supplementary Table 14. Metabolite data shown are the combined results from three independent experiments, each of which included 12 biological replicates each for SLC16A11 and control.

appreciable levels) and profiled ~300 polar and lipid metabolites. Expression of SLC16A11 resulted in substantial increases in triacylglycerol (TAG) levels ($P = 7.6 \times 10^{-12}$), with smaller increases in intracellular diacylglycerols ($P = 7.8 \times 10^{-3}$) and decreases in lysophosphatidylcholine ($P = 2.0 \times 10^{-3}$), cholesterol ester ($P = 9.8 \times 10^{-4}$) and sphingomyelin ($P = 3.9 \times 10^{-3}$) lipids (Fig. 2b, c and Supplementary Tables 13 and 14). As TAG synthesis takes place in the endoplasmic reticulum in the liver[21], these results indicate that SLC16A11 may have a role in hepatic lipid metabolism. We note that serum levels of specific TAGs have been prospectively associated with future risk of type 2 diabetes[22] and accumulation of intracellular lipids has been implicated in insulin resistance in human populations[23,24].

In summary, GWAS in Mexican and other Latin American samples identified a haplotype containing four missense SNPs, all in *SLC16A11*, that is much more common in individuals with Native American ancestry than in other populations. Each haplotype copy is associated with a ~20% increased risk of type 2 diabetes. With these properties, the haplotype would be expected to contribute to the higher burden of type 2 diabetes in Mexican and Latin American populations[25]. The haplotype derives from Neanderthal introgression, providing an example of Neanderthal admixture affecting physiology and disease susceptibility today. Our data suggest the hypothesis for future studies that *SLC16A11* may influence diabetes risk through effects on lipid metabolism in the liver. Our results also indicate that genetic mapping in understudied populations can identify previously undiscovered aspects of disease pathophysiology[1].

Note added in proof: While this paper was in final revision, Hara *et al*. reported[29] a SNP in *SLC16A13* (rs312457) as associated with risk of T2D in an east Asian population with OR = 1.20, $P = 10^{-12}$.

METHODS SUMMARY

DNA samples were prepared using strict quality control procedures and genotyped using the Illumina HumanOmni2.5 array. Stringent sample and SNP quality (including ancestry) filters were applied on the resulting genotypes. After imputation[7,8], SNPs were quality filtered (MAF ≥1% and info score ≥0.6) and association testing was performed via LTSOFT[10] with type 2 diabetes status, BMI, and age modelling liability and adjusting for sex and top two principal components as fixed effect covariates. *P* values were corrected for genomic control ($\lambda_{GC} = 1.046$). Odds ratios (ORs) are from logistic regression in PLINK[26] using BMI, age, sex, and top 2 principal components as covariates. Proportion of Native American ancestry was estimated using ADMIXTURE[27] ($K = 3$) run including unadmixed individuals from several populations.

Odds ratios for young (≤45 years) and older age of onset cases were calculated using logistic regression in each group compared to two randomly selected non-overlapping sets of controls. Significance testing used a *Z*-score calculated from these odds ratios.

Population prevalence was modelled using odds ratio to approximate relative risk in a log-additive effect model[28]. Relative change in population prevalences is reported based on removing a locus with relative risk of 1.20 and the indicated frequency.

Gene expression analyses were performed on data collected using Nanostring and a compendium of publicly available Affymetrix U133 Plus 2.0 microarrays. The subcellular localization of SLC16A11–V5 and metabolic profiling studies were performed after expression of carboxy-terminus, V5-tagged SLC16A11 in HeLa cells. Metabolite values were normalized to the total metabolite signal obtained for each sample. Measurements were obtained in replicate from each of three independent experiments, with data combined after subtracting the mean of the log-transformed values. The Wilcoxon rank sum test was used to test for differences in individual metabolite levels in cells expressing SLC16A11 compared to controls; the Wilcoxon signed rank test was used to assess differences in lipid classes.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 November 2012; accepted 4 November 2013.
Published online 25 December 2013.

1. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nature Rev. Genet.* **11,** 356–366 (2010).
2. Grant, S. F. A. *et al.* Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nature Genet.* **38,** 320–323 (2006).
3. Unoki, H. *et al.* SNPs in *KCNQ1* are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nature Genet.* **40,** 1098–1102 (2008).
4. Yasuda, K. *et al.* Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus. *Nature Genet.* **40,** 1092–1097 (2008).
5. Villalpando, S. *et al.* Prevalence and distribution of type 2 diabetes mellitus in Mexican adult population: a probabilistic survey. *Salud Publica Mex.* **52,** S19–S26 (2010).
6. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
7. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
8. Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H. & Reich, D. Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* **91,** 238–251 (2012).
9. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
10. Zaitlen, N. *et al.* Informed conditioning on clinical covariates increases power in case-control aAssociation Studies. *PLoS Genet.* **8,** e1003032 http://dx.doi.org/10.1371/journal.pgen.1003032 (2012).
11. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genet.* **42,** 579–589 (2010).
12. Halestrap, A. P. The monocarboxylate transporter family—Structure and functional characterization. *IUBMB Life* **64,** 1–9 (2012).
13. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).
14. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
15. Meyer, M. *et al.* A high-coverage genome sequence from an Archaic Denisovan individual. *Science* **338,** 222–226 (2012).
16. Mednikova, M. B. A proximal pedal phalanx of a Paleolithic hominin from denisova cave, Altai. *Archaeol. Ethnol. Anthropol. Eurasia* **39,** 129–138 (2011).
17. Max Planck Institute for Evolutionary Anthropology. A high-quality Neandertal genome sequence. http://www.eva.mpg.de/neandertal/ (2013).
18. Hublin, J. J. The origin of Neandertals. *Proc. Natl Acad. Sci. USA* **106,** 16022–16027 (2009).
19. Halestrap, A. P. & Wilson, M. C. The monocarboxylate transporter family—Role and regulation. *IUBMB Life* **64,** 109–119 (2012).
20. Garcia, C. K., Goldstein, J. L., Pathak, R. K., Anderson, R. G. W. & Brown, M. S. Molecular characterization of a membrane transporter for lactate, pyruvate, and other monocarboxylates: Implications for the Cori cycle. *Cell* **76,** 865–873 (1994).
21. Fu, S., Watkins, S. M. & Hotamisligil, G. S. The role of endoplasmic reticulum in hepatic lipid homeostasis and stress signaling. *Cell Metab.* **15,** 623–634 (2012).
22. Rhee, E. P. *et al.* Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. *J. Clin. Invest.* **121,** 1402–1411 (2011).
23. Savage, D. B. & Semple, R. K. Recent insights into fatty liver, metabolic dyslipidaemia and their links to insulin resistance. *Curr. Opin. Lipidol.* **21,** 329–336 (2010).
24. Samuel, V. T. & Shulman, G. I. Mechanisms for insulin resistance: Common threads and missing links. *Cell* **148,** 852–871 (2012).
25. Florez, J. *et al.* Strong association of socioeconomic status with genetic ancestry in Latinos: implications for admixture studies of type 2 diabetes. *Diabetologia* **52,** 1528–1536 (2009).
26. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
27. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
28. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science.* **273,** 1516–1517 (1996).
29. Hara, K. *et al.* Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum. Mol. Genet.*. http://dx.doi.org/10.1093/hmg/ddt399 (14 August 2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Daly, V. Mootha, E. Lander and K. Estrada for comments on the manuscript, B. Voight, A. Segre, J. Pickrell and the Scientific Advisory Board of the SIGMA Project (especially C. Bustamante) for useful discussions, and A. Subramanian and V. Rusu for assistance with expression analyses. This work was conducted as part of the Slim Initiative for Genomic Medicine, a joint US–Mexico project funded by the Carlos Slim Health Institute. The UNAM/INCMNSZ Diabetes Study was supported by Consejo Nacional de Ciencia y Tecnología grants 138826, 128877, CONACyT- SALUD 2009-01-115250, and a grant from Dirección General de Asuntos del Personal Académico, UNAM, IT 214711. The Diabetes in Mexico Study was supported by Consejo Nacional de Ciencia y Tecnología grant 86867 and by Instituto Carlos Slim de la Salud, A.C. The Mexico City Diabetes Study was supported by National Institutes of Health (NIH) grant R01HL24799 and by the Consejo Nacional de Ciencia y Tenologia grants 2092, M9303, F677-M9407, 251M and 2005-C01-14502, SALUD 2010-2-151165. The Multiethnic Cohort was supported by NIH grants CA164973, CA054281 and CA063464. The Singapore Chinese Health Study was funded by the National Medical Research Council of Singapore under its individual research grant scheme and by NIH grants R01 CA55069, R35 CA53890, R01 CA80205 and R01 CA144034. The Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) project was supported by NIH grant U01DK085526. The San Antonio Mexican American Family Studies (SAMAFS) were supported by R01 DK042273, R01 DK047482, R01 DK053889, R01 DK057295, P01 HL045522 and a Veterans Administration Epidemiologic grant to R.A.D. A.L.W. was

supported by National Institutes of Health Ruth L. Kirschstein National Research Service Award number F32 HG005944.

Author Contributions See the author list for details of author contributions.

Author Information Genotype data have been deposited in dbGaP under accession number phs000683.v1.p1. Microarray data used in the '55k screen' is publicly available through the NCBI Gene Expression Omnibus and the Cancer Cell Line Encyclopedia. A list of sample identities and accession numbers are available in the Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A. (altshuler@molbio.mgh.harvard.edu) or T.T.L. (mttusie@gmail.com).

The SIGMA Type 2 Diabetes Genetics Consortium

Writing team: Amy L. Williams[1,2], Suzanne B. R. Jacobs[1], Hortensia Moreno-Macías[3], Alicia Huerta-Chagoya[4,5], Claire Churchhouse[1], Carla Márquez-Luna[6], Humberto García-Ortíz[6], María José Gómez-Vázquez[4,7], Noël P. Burtt[1], Carlos A. Aguilar-Salinas[4], Clicerio González-Villalpando[8], Jose C. Florez[1,9,10], Lorena Orozco[6], Christopher A. Haiman[11], Teresa Tusié-Luna[4,5], David Altshuler[1,2,9,10,12,13,14]

Analysis team: Amy L. Williams[1,2], Carla Márquez-Luna[6], Alicia Huerta-Chagoya[4,5], Stephan Ripke[1,15], María José Gómez-Vázquez[4,7], Alisa K. Manning[1], Hortensia Moreno-Macías[3], Humberto García-Ortíz[6], Benjamin Neale[1,15], Noël P. Burtt[1], Carlos A. Aguilar-Salinas[4], David Reich[1,2], Daniel O. Stram[11], Juan Carlos Fernández-López[6], Sandra Romero-Hidalgo[6], David Altshuler[1,2,9,10,12,13,14], Jose C. Florez[1,9,10], Teresa Tusié-Luna[4,5], Nick Patterson[1], Christopher A. Haiman[11]

Clinical research, study design and metabolic phenotyping: Diabetes in Mexico Study Irma Aguilar-Delfín[6], Angélica Martínez-Hernández[6], Federico Centeno-Cruz[6], Elvia Mendoza-Caamal[6], Cristina Revilla-Monsalve[16], Sergio Islas-Andrade[16], Emilio Córdova[6], Eunice Rodríguez-Arellano[17], Xavier Soberón[6], Lorena Orozco[6]; **Massachusetts General Hospital** Jose C. Florez[1,9,10]; **Mexico City Diabetes Study** Clicerio González-Villalpando[8], María Elena González-Villalpando[8]; **Multiethnic Cohort** Christopher A. Haiman[11], Brian E. Henderson[11], Kristine Monroe[11], Lynne Wilkens[18], Laurence N. Kolonel[18], Loic Le Marchand[18]; **UNAM/INCMNSZ Diabetes Study** Laura Riba[5], María Luisa Ordóñez-Sánchez[4], Rosario Rodríguez-Guillén[4], Ivette Cruz-Bautista[4], Maribel Rodríguez-Torres[4], Linda Liliana Muñoz-Hernández[4], Tamara Sáenz[4], Donají Gómez[4], Ulices Alvirde[4]

Sample quality control and whole-genome genotyping: Noël P. Burtt[1], Robert C. Onofrio[19], Wendy M. Brodeur[19], Diane Gage[19], Jacquelyn Murphy[1], Jennifer Franklin[19], Scott Mahan[19], Kristin Ardlie[19], Andrew T. Crenshaw[19], Wendy Winckler[19]

Neanderthal analysis team: Kay Prüfer[20], Michael V. Shunkov[21], Susanna Sawyer[20], Udo Stenzel[20], Janet Kelso[20], Monkol Lek[1,15], Sriram Sankararaman[1,2], Amy L. Williams[1,2], Nick Patterson[1], Daniel G. MacArthur[1,15], David Reich[1,2], Anatoli P. Derevianko[21], Svante Pääbo[20]

Functional analysis and metabolite profiling: Suzanne B. R. Jacobs[1], Claire Churchhouse[1], Shuba Gopal[22], James A. Grammatikos[22], Ian C. Smith[23], Kevin H. Bullock[22], Amy A. Deik[22], Amanda L. Souza[22], Kerry A. Pierce[22], Clary B. Clish[22], David Altshuler[1,2,9,10,12,13,14]

Replication genotyping and analysis: Broad Institute of Harvard and MIT Timothy Fennell[19], Yossi Farjoun[19], Broad Genomics Platform*, Stacey Gabriel[19]; **Singapore Chinese Health Study** Daniel O. Stram[11], Myron D. Gross[24], Mark A. Pereira[24], Mark Seielstad[25], Woon-Puay Koh[26,27], E-Shyong Tai[26,27,28]; **T2D-GENES Consortium** Jason Flannick[1,9], Pierre Fontanillas[1], Andrew Morris[29], Tanya M. Teslovich[30], Noël P. Burtt[1], Gil Atzmon[31], John Blangero[32], Donald W. Bowden[33], John Chambers[34,35,36], Yoon Shin Cho[37], Ravindranath Duggirala[32], Benjamin Glaser[38,39], Craig Hanis[40], Jaspal Kooner[35,36,41], Markku Laakso[42], Jong-Young Lee[43], E-Shyong Tai[26,27,28], Yik Ying Teo[44,45,46,47,48], James G. Wilson[49], the T2D-GENES Consortium*; **Multiethnic Cohort** Christopher A. Haiman[11], Brian E. Henderson[11], Kristine Monroe[11], Lynne Wilkens[18], Laurence N. Kolonel[18], Loic Le Marchand[18]; **Texas Biomedical Research Institute and University of Texas Health Science Center at San Antonio** Sobha Puppala[32], Vidya S. Farook[32], Farook Thameem[50], Hanna E. Abboud[50], Ralph A. DeFronzo[51], Christopher P. Jenkinson[51], Donna M. Lehman[52], Joanne E. Curran[32], John Blangero[32], Ravindranath Duggirala[32]

Scientific and project management: Noël P. Burtt[1], Maria L. Cortes[53]

Steering committee: David Altshuler[1,2,9,10,12,13,14], Jose C. Florez[1,9,10], Christopher A. Haiman[11], Brian E. Henderson[11], Carlos A. Aguilar-Salinas[4], Clicerio González-Villalpando[8], Lorena Orozco[6] & Teresa Tusié-Luna[4,5]

[1]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Universidad Autonoma Metropolitana, Tlalpan 14387, Mexico City, Mexico. [4]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Sección XVI, Tlalpan, 14000 Mexico City, Mexico. [5]Instituto de Investigaciones Biomédicas, UNAM. Unidad de Biología Molecular y Medicina Genómica, UNAM/INCMNSZ, Coyoacán, 04510 Mexico City, Mexico. [6]Instituto Nacional de Medicina Genómica, Tlalpan, 14610 Mexico City, Mexico. [7]Universidad Autónoma de Nuevo León, San Nicolás de los Garza, Nuevo León 66451, México. [8]Centro de Estudios en Diabetes, Unidad de Investigacion en Diabetes y Riesgo Cardiovascular, Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, 01120 Mexico City, Mexico. [9]Center for Human Genetic Research and Diabetes Research Center (Diabetes Unit), Massachusetts General Hospital, Boston 02114, Massachusetts, USA. [10]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [11]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90089, USA. [12]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [13]Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts 02114, USA. [14]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [15]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [16]Unidad de Investigación Médica en Enfermedades Metabólicas, Instituto Mexicano del Seguro Social SXXI, Cuauhtémoc, 06720 Mexico City, Mexico. [17]Instituto de Seguridad y Servicios Sociales para los Trabajadores del Estado, Álvaro Obregón, 01030 Mexico City, Mexico. [18]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii 96813, USA. [19]The Genomics Platform, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [20]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany. [21]Palaeolithic Department, Institute of Archaeology and Ethnography, Russian Academy of Sciences, Siberian Branch, 630090 Novosibirsk, Russia. [22]The Metabolite Profiling Platform, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [23]Cancer Biology Program, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [24]University of Minnesota, Minneapolis, Minnesota 55455, USA. [25]University of California San Francisco, San Francisco, California 94143, USA. [26]Duke National University of Singapore Graduate Medical School, Singapore 169857, Singapore. [27]Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117597, Singapore. [28]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore. [29]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [30]Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA. [31]Department of Medicine, Department of Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA. [32]Department of Genetics, Texas Biomedical Research Institute, San Antonio, Texas 78227, USA. [33]Center for Genomics and Personalized Medicine Research, Center for Diabetes Research, Department of Biochemistry, Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, North Carolina 27157, USA. [34]Department of Epidemiology and Biostatistics, Imperial College London, London SW7 2AZ, UK. [35]Imperial College Healthcare NHS Trust, London W2 1NY, UK. [36]Ealing Hospital National Health Service (NHS) Trust, Middlesex UB1 3HW, UK. [37]Department of Biomedical Science, Hallym University, Chuncheon, Gangwon-do, 200-702 South Korea. [38]Endocrinology and Metabolism Service, Hadassah-Hebrew University Medical School, Jerusalem 91120, Israel. [39]Israel Diabetes Research Group (IDRG), Diabetes Unit, The E. Wolfson Medical Center, Holon 58100, Israel. [40]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. [41]National Heart and Lung Institute (NHLI), Imperial College London, Hammersmith Hospital, London W12 0HS, UK. [42]Department of Medicine, University of Eastern Finland, Kuopio Campus and Kuopio University Hospital, FI-70211 Kuopio, Finland. [43]Center for Genome Science, Korea National Institute of Health, Osong Health Technology Administration Complex, Chungcheongbuk-do 363-951, South Korea. [44]Department of Epidemiology and Public Health, National University of Singapore, Singapore 117597, Singapore. [45]Centre for Molecular Epidemiology, National University of Singapore, Singapore 117456, Singapore. [46]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore. [47]Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456, Singapore. [48]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore. [49]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi 39216, USA. [50]Division of Nephrology, Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, USA. [51]Division of Diabetes, Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, USA. [52]Division of Clinical Epidemiology, Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, USA. [53]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. *Lists of participants and their affiliations appear in the Supplementary Information.

doi:10.1038/nature12720

Structure of the SecY channel during initiation of protein translocation

Eunyong Park[1], Jean-François Ménétret[2], James C. Gumbart[3], Steven J. Ludtke[4], Weikai Li[1], Andrew Whynot[1], Tom A. Rapoport[1] & Christopher W. Akey[2]

Many secretory proteins are targeted by signal sequences to a protein-conducting channel, formed by prokaryotic SecY or eukaryotic Sec61 complexes, and are translocated across the membrane during their synthesis[1,2]. Crystal structures of the inactive channel show that the SecY subunit of the heterotrimeric complex consists of two halves that form an hourglass-shaped pore with a constriction in the middle of the membrane and a lateral gate that faces the lipid phase[3–5]. The closed channel has an empty cytoplasmic funnel and an extracellular funnel that is filled with a small helical domain, called the plug. During initiation of translocation, a ribosome–nascent chain complex binds to the SecY (or Sec61) complex, resulting in insertion of the nascent chain. However, the mechanism of channel opening during translocation is unclear. Here we have addressed this question by determining structures of inactive and active ribosome–channel complexes with cryo-electron microscopy. Non-translating ribosome–SecY channel complexes derived from *Methanocaldococcus jannaschii* or *Escherichia coli* show the channel in its closed state, and indicate that ribosome binding per se causes only minor changes. The structure of an active *E. coli* ribosome–channel complex demonstrates that the nascent chain opens the channel, causing mostly rigid body movements of the amino- and carboxy-terminal halves of SecY. In this early translocation intermediate, the polypeptide inserts as a loop into the SecY channel with the hydrophobic signal sequence intercalated into the open lateral gate. The nascent chain also forms a loop on the cytoplasmic surface of SecY rather than entering the channel directly.

Opening of the SecY channel during initiation of translocation involves two events: binding of the ribosome and insertion of the nascent chain. To analyse how ribosome binding per se affects the structure of a translocation channel, we first determined the structure of complexes lacking a nascent chain. Initial experiments were performed with complexes from *M. jannaschii*, because this allows a direct comparison with a crystal structure of SecY[3]. Purified *M. jannaschii* ribosomes were incubated with an excess of SecY complex, and complexes were imaged by cryo-electron microscopy. A total of ~37,000 particles were analysed, resulting in an electron density map with a resolution of 9.0 Å for the ribosome and ~12.7 Å for the channel (Supplementary Table 1).

A ribosome model from *Pyrococcus furiosus*[6], a species related to *M. jannaschii*, was fit into the density map, allowing the identification of essentially all RNA helices and many helical features of ribosomal proteins (Fig. 1a and Supplementary Fig. 1). A crystal structure of the *M. jannaschii* SecY complex could be docked into density for the SecY channel (Fig. 1b and Supplementary Fig. 2), and molecular dynamics flexible fitting (MDFF)[7] resulted in only small changes (Fig. 1c). All transmembrane segments (TMs), including the 10 TMs of SecY, and the single TMs of the SecE and Secβ subunits, could be accounted for in the map. Several TM helices and the extracellular loop between TMs 5 and 6 were partially resolved (Supplementary Fig. 3). A comparison with the crystal structure shows that, with the exception of some adjustments in the cytoplasmic helix of SecE, membrane-embedded domains remained essentially unaltered (Fig. 1c). As observed previously with other species[8–11], loops between TMs 6 and 7 (6/7 loop) and TMs 8 and 9 (8/9 loop) of SecY, as well as the cytoplasmic helix of SecE (Fig. 1b), all interact with components of the large ribosomal subunit at the tunnel exit (Supplementary Fig. 4a–c). These interactions do not induce major structural changes in the SecY channel and leave the lateral gate closed.

Next we determined the structure of a non-translating ribosome–channel complex from *E. coli*, with a larger data set than used previously[8]. A total of ~39,000 particles were analysed, resulting in a density map with a resolution of ~9.5 Å for the ribosome and ~14 Å for the channel (Supplementary Table 1). Models for ribosomal subunits[11,12] were docked into the density map (Fig. 1d) and all RNA helices were visible, as well as some partially resolved helices of ribosomal proteins (Supplementary Fig. 5). Because there is no crystal structure of the *E. coli* SecY complex, we generated a homology model on the basis of crystal structures of *Thermus thermophilus* and *Thermotoga maritima* complexes[4,13] (Supplementary Figs 6 and 7). This model was subjected to MDFF using the entire density map of the ribosomal large subunit and channel as a restraint. This resulted in movements of cytoplasmic loops, whereas membrane-embedded domains remained essentially unchanged (Supplementary Fig. 8). Many features of the channel are clearly visible in a segmented map (Fig. 1e and Supplementary Figs 9 and 10), including cytoplasmic loops of SecY, two helices of SecE, two TMs of SecG (the bacterial equivalent of archaeal Secβ) and some partially resolved TMs of SecY. Connections between the channel and ribosome were similar to those in the *M. jannaschii* complex, with the exception of the longer 6/7 loop of SecY, which is repositioned between RNA helices 6 and 7 (Supplementary Fig. 4d–f). Importantly, the ribosome alone does not induce major changes in the channel structure, so the lateral gate remains closed (Fig. 1f).

To determine the structure of an active *E. coli* ribosome–channel complex, we used a new strategy. Previous attempts to obtain a structure of an active translocation channel showed that a translating ribosome was bound to the channel, but there was little biochemical evidence that a nascent chain was inserted in the channel and no clear electron density was visible for the polypeptide[10,11]. These studies used small amounts of ribosome–nascent chain complexes (RNCs) that were formed *in vitro* and subsequently added to purified channels. To obtain a more physiological sample, we generated an early translocation intermediate of a secretory protein in living *E. coli* cells by expressing a polypeptide with 100 amino acids from an inducible promoter[14,15]. The polypeptide has an N-terminal signal sequence derived from DsbA, which targets the protein to the co-translational translocation pathway[16], and a C-terminal SecM-stalling sequence, which arrests translation of the ribosome[17] (Fig. 2a). We also expressed the endoribonuclease MazF from an inducible promoter to cleave messenger RNA between ribosomes, which results in the depletion of nascent chains associated with non-stalled ribosomes[18]. To generate a stable complex between the SecM-stalled

[1]Department of Cell Biology and Howard Hughes Medical Institute, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. [2]Department of Physiology and Biophysics, Boston University School of Medicine, 700 Albany Street, Boston, Massachusetts 02118-2526, USA. [3]School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA. [4]National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA.

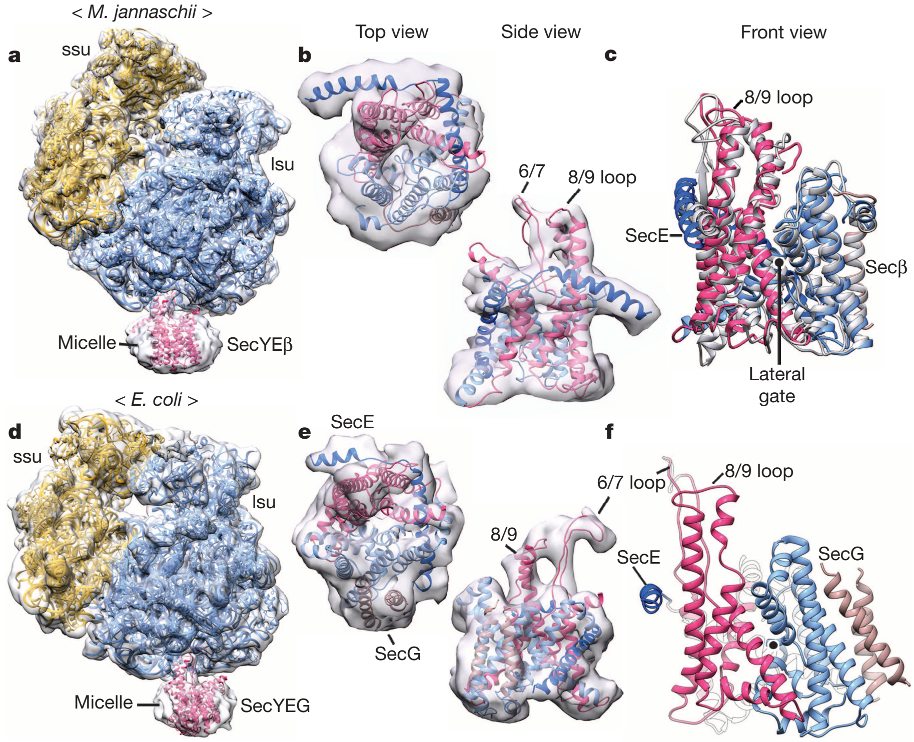


Figure 1 | **Structures of non-translating ribosome–channel complexes.** **a**, Density map for the *M. jannaschii* complex. Models for ribosomal RNA and proteins of the small and large ribosomal subunits (ssu and lsu; in gold and blue, respectively) and of the SecY complex (in red) were docked into the map. **b**, Fit of the *M. jannaschii* SecY complex into the segmented density map, as viewed from the cytoplasm (top view) and from the side. The N- and C-terminal halves of SecY are in light blue and red, respectively. SecE is in dark blue and Secβ in brown. **c**, Comparison between the crystal structure of an *M. jannaschii* SecY complex (grey) and the electron microscopy structure (in colour), as viewed facing the lateral gate (front view). **d**, **e**, As in **a** and **b**, but for the *E. coli* complex. SecG, the bacterial equivalent of Secβ, is in brown. **f**, A model for the *E. coli* channel in a front view.

RNC and the channel, we used disulphide crosslinking. The nascent chain contained a cysteine at position 19 of the signal sequence, which can be crosslinked to a cysteine at position 68 in the SecY plug[14]. Disulphide bond formation was achieved by adding an oxidant to the *E. coli* culture, resulting in 70% of nascent chains being linked to SecY.

To purify the RNC–channel complex, we replaced the endogenous ribosomal protein L12 with a *Strep*-tagged version, allowing the enrichment of ribosomes on a *Strep*-Tactin column. This purification step was performed at high salt concentration to remove SecY complexes lacking a nascent chain (Supplementary Fig. 11a). A second purification step exploited a His-tag inserted into a fusion between SecE and SecG, and allowed the enrichment of channel-containing complexes by Co^{2+}-affinity chromatography. Finally, the sample was subjected to gel filtration. The purified RNC–channel complex eluted as a homogeneous peak at the position of monosomes (Supplementary Fig. 11b). On a Coomassie-stained SDS gel, the SecY–nascent chain–transfer RNA species was the only major band besides those from ribosomal proteins (Fig. 2b, lane 1). As expected, the band disappeared when the sample was treated with a reducing agent to remove the disulphide bridge or with RNase A to degrade the tRNA (Fig. 2b, lanes 2 and 3). We found that the previous protocol of adding purified RNCs to SecY complex, either in detergent or in nanodiscs[10,11], resulted in inefficient insertion of the nascent chain into the channel (Supplementary Fig. 12). Also, when RNC–channel complexes were generated *in vivo* and crosslinked after purification, crosslinks between different nascent chain molecules and between the nascent chain and unidentified proteins were observed (Supplementary Fig. 13). Hence, crosslinking *in vivo* is required to maintain the nascent chain in the channel.

Purified RNC–channel complexes were frozen over holes on electron microscopy grids, as the channel was lost when complexes were placed on a carbon film. A total of ~167,000 individual particles were used, of which ~50% contained the channel. Additional sorting for the best signal-to-noise ratio identified ~53,000 particles for structure determination and resulted in a density map at ~10 Å resolution for the ribosome and ~11 Å for the channel (Fig. 3a and Supplementary Table 1). Ribosomal RNAs and proteins were clearly visible in the density map (Supplementary Fig. 14), along with aminoacyl (A-site) and peptidyl (P-site) tRNAs, as expected for a SecM-stalled ribosome[19] (Supplementary Fig. 15a). Moreover, there was density for mRNA underneath the anticodon regions of tRNAs (Supplementary Fig. 15b). We

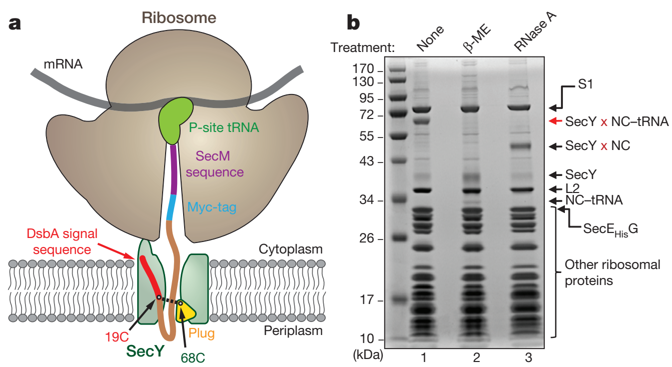


Figure 2 | **Purification of a RNC–channel complex.** **a**, The complex was generated in living *E. coli* cells by expressing a nascent chain (NC) of 100 amino acids with a signal sequence and SecM-stalling sequence. The nascent chain also contains a Myc-tag. A cysteine at position 19 of the nascent chain (19C) was disulphide-crosslinked to a cysteine in the plug of SecY (68C). **b**, Coomassie-stained SDS-gel of the ribosome–NC (RNC)–channel complex (lane 1). The red arrow indicates the crosslinked product of SecY and the NC–tRNA adduct. This band disappears after treatment with β-mercaptoethanol (β-ME) or RNaseA (lanes 2 and 3). Ribosomal proteins (including S1) and the fusion between SecE and SecG are indicated.

also observed density for ribosomal protein S1 that was more extensive than seen before[20] (Fig. 3a and Supplementary Fig. 15c–e).

To generate a model for the active channel, we created an *E. coli* homology model on the basis of a crystal structure of the SecY complex from *P. furiosus*[5], which has the most open lateral gate among known crystal structures (Supplementary Fig. 16), and used MDFF to adjust the model to the experimental density map. The 6/7 loop and TM9 of SecY were well resolved (Fig. 3b), and ribosomal components interacting with the channel were the same as with the non-translating complex. The cytoplasmic helix of SecE and TM10 of SecY were clearly visible, and there was good density for SecG (Supplementary Figs 17 and 18). In addition, many TMs were partially resolved, with only occasional density breaks in the helices. Density for the nascent chain was clearly identifiable without segmentation of the density map. Specifically, additional density for a helix was visible in the cytoplasmic part of the lateral gate (see below), explaining why a channel with a fully open lateral gate could be fit into the density map. In fact, the lateral gate is more open than in the *P. furiosus* crystal structure[5] (Supplementary Table 2). Calculated cross correlation coefficients showed that the model for the open SecY channel is a better fit in the density map than the model for the closed channel (Supplementary Table 1).

The modelled conformational change of the *E. coli* channel is supported by the fact that the conversion from a closed to an open channel involves mostly rigid body movements of the N- and C-terminal halves of SecY (Supplementary Fig. 19). To open the lateral gate, the N-terminal half of SecY undergoes a large rotation and tilt, whereas the C-terminal half moves less in the opposite direction (Fig. 3c; see also Supplementary Video 1). SecE undergoes a tilting motion to accommodate movements of SecY, and SecG moves with the N-terminal half of SecY. These conformational changes would maintain the hydrophobic belt of the SecY complex within the lipid environment. In addition to rigid body movements, there are changes in the 5/6 loop that connect the two halves of SecY to accommodate the large opening motion. There are also movements in TM8 and the lower part of TM7. One particularly large change occurs in the upper part of TM8 (helix 8b), which is displaced towards the membrane surface (Fig. 3d). The 6/7 loop and TM9, as well as preceding loop residues, including a conserved arginine (Arg 357), do not move appreciably (Fig. 3d), consistent with their role in tethering the channel to the ribosome. The plug domain moves only a small distance, probably because it is restrained by the disulphide bridge to the signal sequence. However, the plug does not have to move much to allow translocation[21]. When viewed from the cytoplasmic side, these conformational changes open a pore adjacent to the lateral gate (Fig. 3e; see also Supplementary Video 2). Overall, the changes are more pronounced than seen previously[10,11].

Density for the nascent chain was seen inside the ribosomal tunnel, on the cytoplasmic surface of the SecY complex, inside the channel, and on its periplasmic side (Fig. 4a and Supplementary Fig. 20). On the

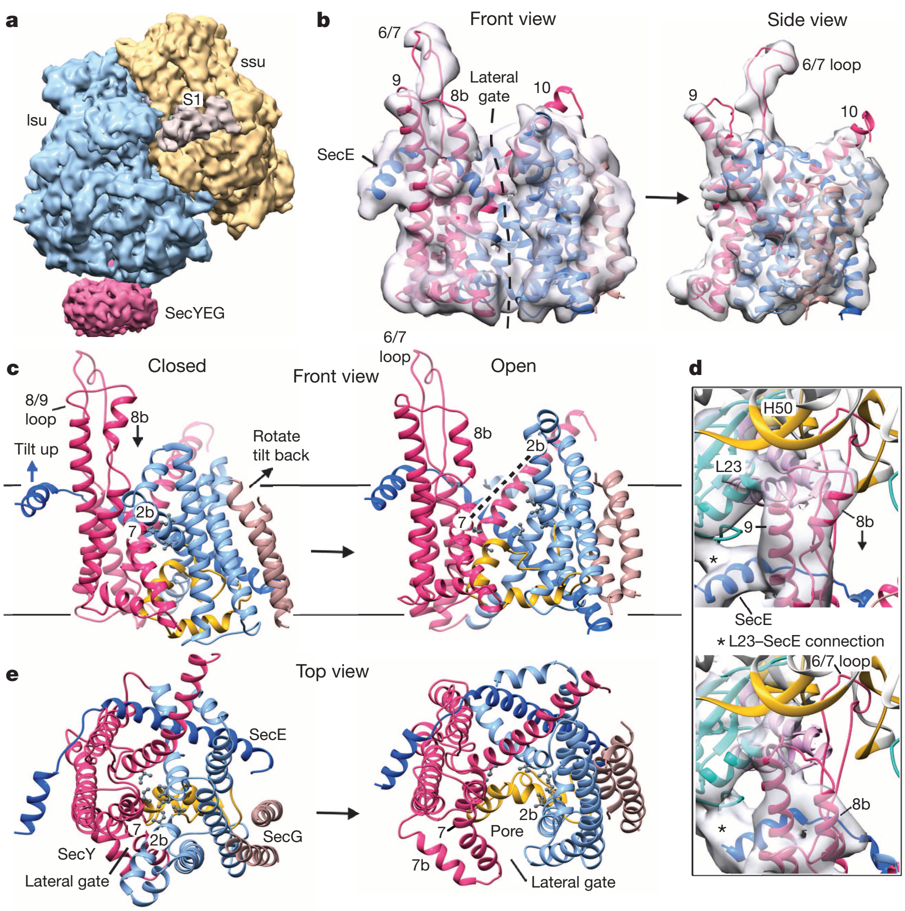


Figure 3 | Structure of the active SecY channel. **a**, Structure of the *E. coli* RNC–SecY channel complex, with large and small ribosomal subunits in blue and gold, respectively, the SecY complex in red, and ribosomal protein S1 in tan. **b**, Front (left) and side (right) views of the channel fit into the segmented density map (grey). The nascent chain was omitted for clarity. The N-terminal half of SecY is in light blue, the C-terminal half in red, SecE in dark blue and SecG in brown. **c**, Comparison of front views of the closed (left) and open (right) *E. coli* SecY channels, with the approximate position of the membrane indicated by solid horizontal lines. The N-terminal half of SecY is in light blue, the C-terminal half in red, SecE in dark blue, SecG in brown, and the plug in yellow. Some movements during channel opening are indicated, such as the rotation and tilting of the N-terminal half of SecY, the tilting of SecE, and the movement of helix 8b. Labels for helices 2b and 7 are placed at the same position in the closed and open channel. Pore residues forming the constriction in the closed channel are indicated with grey balls and sticks. **d**, Connections of the ribosome with the 8/9 loop of SecY and the cytoplasmic helix of SecE in the closed and open channels (top and bottom panels, respectively). Note the large movement of helix 8b towards the membrane. **e**, As in **c**, but viewed from the top.

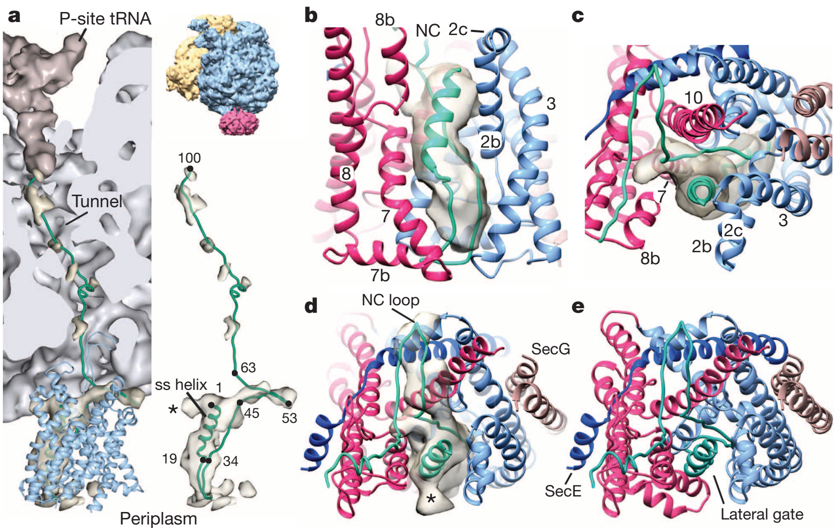


Figure 4 | Path of the nascent chain. **a**, Density (in light gold) and model (green line) for the nascent chain in the RNC–channel complex. The P-site tRNA is in brown, the ribosome in grey, and the channel in blue. The top right panel shows the entire RNC–channel complex from the same viewing angle. The bottom right panel shows the density and model for the nascent chain, with ribosome and channel omitted. The asterisk indicates density for an alternative orientation of the nascent chain loop on the cytoplasmic side of the channel (see also **d**). **b**, Side view of the signal sequence (ss) helix in the lateral gate. Density for the nascent chain on the cytoplasmic surface was removed for clarity. **c**, As in **b**, but viewed from the top along the axis of the signal sequence helix. **d**, As in **c**, but from a slightly different angle of view with nascent chain density on the cytoplasmic surface included. **e**, As in **d**, but without the density map.

basis of biochemical data[14], an approximate model for the nascent chain in the RNC–channel complex was built into the density. The last ~40 amino acids are located inside the ribosome, as cysteines introduced into this segment are inaccessible to a bulky modification reagent. In addition, cysteines at positions 19–34 are most favoured to form a disulphide bridge with a cysteine in the plug. Finally, the position of the end of the signal sequence in our structure is constrained by the disulphide crosslink between position 19 of the nascent chain and position 68 of the plug.

The resulting model shows that the hydrophobic core of the signal sequence forms a helix in the lateral gate (residues 1–15) (Fig. 4b–d and Supplementary Fig. 21), consistent with crosslinking data obtained with the yeast Sec61 complex[22]. The signal sequence helix is contacted by TM2b, helix 8b and TM7 of SecY (Fig. 4b). In a lipid bilayer, much of the signal sequence, including parts that follow the hydrophobic region, would be exposed to the hydrocarbon chains of phospholipids, again in agreement with crosslinking experiments[22]. Additional density below and adjacent to the signal sequence helix can account for the other side of the nascent chain loop. The pore through which the mature region of the nascent chain would move into the extracellular funnel is not exactly in the centre of the channel, but the translocating polypeptide may still be surrounded by pore ring residues that form a constriction in the closed channel (Supplementary Video 2). Crosslinking to the nascent chain may restrain the plug, keeping it in the centre of the channel. However, there is still room for the nascent chain to form a loop in the pore.

We modelled density on the cytoplasmic surface of the channel as a loop that extends parallel to the surface and towards the back of the channel (residues ~45–63) (Fig. 4a, e). This part of the nascent chain lies in a V-shaped groove, which is framed by the base of the 6/7 loop and TM10 of SecY (Supplementary Fig. 22 and Supplementary Video 3). However, the nascent chain may adopt an alternative orientation with a loop that extends above the lateral gate (marked with an asterisk in Fig. 4a, d). The nascent chain may also slide up and down the axis of the channel to some extent, as there is density on the periplasmic side that is not fully accounted for in our model.

In summary, our structures show that ribosome binding alone does not induce major changes in the SecY channel, although it may cause transient opening[23]. Rather, stable opening of the channel requires loop insertion of the nascent chain[24]. As predicted[3,22], the hydrophobic part of the signal sequence forms a helix that occupies the open lateral gate. The signal sequence would thus become part of the channel wall, thereby increasing the size of the pore through which the polypeptide moves across the membrane. At later stages of translocation, the signal sequence is cleaved from the nascent chain and released from the lateral gate, which may result in a narrower pore. It is also possible that the signal sequence leaves the lateral gate before cleavage. This hypothesis would be consistent with a two-dimensional crystal structure of the SecY complex that showed a synthetic signal peptide bound to the outside of an essentially closed channel[25].

Our results also indicate that most nascent chains form a loop on the cytoplasmic surface of SecY, rather than adopting a fully extended conformation between the ribosome and channel. Although the observed looping of the nascent chain at the cytoplasmic surface of the channel needs to be confirmed with other substrates, it seems possible that a pulling force or ratcheting mechanism[26,27] may be required to achieve efficient translocation. SecDF could use a proton gradient across the membrane together with movements of a periplasmic domain to pull on the nascent chain[28]. In addition, polypeptide chain folding or the binding of periplasmic chaperones may help to move the polypeptide chain across the membrane.

METHODS SUMMARY

The purification of *E. coli* 70S ribosomes and *M. jannaschii* and *E. coli* SecY complexes were each described previously[3,8]. *M. jannaschii* 70S ribosomes were purified by sucrose gradient centrifugation, dissociated into 50S and 30S subunits, and re-associated. Non-translocating ribosome–SecY complexes were reconstituted by mixing ribosomes with a five- to eightfold molar excess of the SecY channels in *n*-dodecyl-β-D-maltoside (DDM). An RNC–SecY complex was generated in *E. coli* cells by expressing a SecM-stalled nascent chain under the arabinose promoter[14]. In addition, MazF endoribonuclease[18] was expressed from a Tet promoter. After forming a disulphide bridge between the nascent chain and SecY by addition of 5,5′-dithiobis-(2-nitrobenzoic acid), RNC–SecY complexes were solubilized in DDM, and purified by tandem affinity chromatography using a *Strep*-tag on the ribosomal protein L12 and a His-tag on a fusion of SecE and SecG. Complexes were further purified by size-exclusion chromatography on Superose 6. Samples for cryo-electron microscopy were applied to holey grids or to grids with a continuous carbon film and vitrified. Images were collected at 160 and 200 kV on a Tecnai FEG 20 (FEI) with nominal magnifications of ×40,000 and ×52,000, using a TVIPS 4096 × 4096 charge-coupled device or film. Image processing and single-particle analysis were done with EMAN software[29]. Molecular docking was carried out with Chimera[30] and MDFF[7].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 June; accepted 4 October 2013.
Published online 23 October 2013.

1. Park, E. & Rapoport, T. A. Mechanisms of Sec61/SecY-mediated protein translocation across membranes. *Annu. Rev. Biophys.* **41,** 21–40 (2012).
2. Shao, S. & Hegde, R. S. Membrane protein insertion at the endoplasmic reticulum. *Annu. Rev. Cell Dev. Biol.* **27,** 25–56 (2011).
3. Van den Berg, B. *et al.* X-ray structure of a protein-conducting channel. *Nature* **427,** 36–44 (2004).
4. Tsukazaki, T. *et al.* Conformational transition of Sec machinery inferred from bacterial SecYE structures. *Nature* **455,** 988–991 (2008).

5. Egea, P. F. & Stroud, R. M. Lateral opening of a translocon upon entry of protein suggests the mechanism of insertion into membranes. *Proc. Natl Acad. Sci. USA* **107,** 17182–17187 (2010).
6. Armache, J. P. *et al.* Promiscuous behaviour of archaeal ribosomal proteins: implications for eukaryotic ribosome evolution. *Nucleic Acids Res.* **41,** 1284–1293 (2013).
7. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16,** 673–683 (2008).
8. Ménétret, J. F. *et al.* Ribosome binding of a single copy of the SecY complex: implications for protein translocation. *Mol. Cell* **28,** 1083–1092 (2007).
9. Ménétret, J. F. *et al.* Single copies of Sec61 and TRAP associate with a nontranslating mammalian ribosome. *Structure* **16,** 1126–1137 (2008).
10. Becker, T. *et al.* Structure of monomeric yeast and mammalian Sec61 complexes interacting with the translating ribosome. *Science* **326,** 1369–1373 (2009).
11. Frauenfeld, J. *et al.* Cryo-EM structure of the ribosome–SecYE complex in the membrane environment. *Nature Struct. Mol. Biol.* **18,** 614–621 (2011).
12. Berk, V., Zhang, W., Pai, R. D. & Cate, J. H. D. Structural basis for mRNA and tRNA positioning on the ribosome. *Proc. Natl Acad. Sci. USA* **103,** 15830–15834 (2006).
13. Zimmer, J., Nam, Y. & Rapoport, T. A. Structure of a complex of the ATPase SecA and the protein-translocation channel. *Nature* **455,** 936–943 (2008).
14. Park, E. & Rapoport, T. A. Preserving the membrane barrier for small molecules during bacterial protein translocation. *Nature* **473,** 239–242 (2011).
15. Park, E. & Rapoport, T. A. Bacterial protein translocation requires only one copy of the SecY complex *in vivo*. *J. Cell Biol.* **198,** 881–893 (2012).
16. Schierle, C. F. *et al.* The DsbA signal sequence directs efficient, cotranslational export of passenger proteins to the *Escherichia coli* periplasm via the signal recognition particle pathway. *J. Bacteriol.* **185,** 5706–5713 (2003).
17. Nakatogawa, H. & Ito, K. The ribosomal exit tunnel functions as a discriminating gate. *Cell* **108,** 629–636 (2002).
18. Zhang, Y. *et al.* MazF cleaves cellular mRNAs specifically at ACA to block protein synthesis in *Escherichia coli*. *Mol. Cell* **12,** 913–923 (2003).
19. Muto, H., Nakatogawa, H. & Ito, K. Genetically encoded but nonpolypeptide prolyl-tRNA functions in the A site for SecM-mediated ribosomal stall. *Mol. Cell* **22,** 545–552 (2006).
20. Sengupta, J., Agrawal, R. K. & Frank, J. Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. *Proc. Natl Acad. Sci. USA* **98,** 11991–11996 (2001).
21. Lycklama a Nijeholt, J. A., Bulacu, M., Marrink, S. J. & Driessen, A. J. Immobilization of the plug domain inside the SecY channel allows unrestricted protein translocation. *J. Biol. Chem.* **285,** 23747–23754 (2010).
22. Plath, K., Mothes, W., Wilkinson, B. M., Stirling, C. J. & Rapoport, T. A. Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell* **94,** 795–807 (1998).
23. Knyazev, D. G. *et al.* The bacterial translocon SecYEG opens upon ribosome binding. *J. Biol. Chem.* **288,** 17941–17946 (2013).
24. Shaw, A. S., Rottier, P. J. & Rose, J. K. Evidence for the loop model of signal-sequence insertion into the endoplasmic reticulum. *Proc. Natl Acad. Sci. USA* **85,** 7592–7596 (1988).
25. Hizlan, D. *et al.* Structure of the SecY complex unlocked by a preprotein mimic. *Cell Rep* **1,** 21–28 (2012).
26. Matlack, K. E., Misselwitz, B., Plath, K. & Rapoport, T. A. BiP acts as a molecular ratchet during posttranslational transport of prepro-α factor across the ER membrane. *Cell* **97,** 553–564 (1999).
27. Nicchitta, C. V. & Blobel, G. Lumenal proteins of the mammalian endoplasmic reticulum are required to complete protein translocation. *Cell* **73,** 989–998 (1993).
28. Tsukazaki, T. *et al.* Structure and function of a membrane component SecDF that enhances protein export. *Nature* **474,** 235–238 (2011).
29. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).
30. Pettersen, E. F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank K. Matlack and T. Guettler for reading the manuscript. This work was supported by National Institutes of Health grants GM067887 to J.C.G., GM080139 to S.J.L., GM052586 to T.A.R. and GM45377 to C.W.A. T.A.R. is a Howard Hughes Institute investigator.

Author Contributions E.P. designed and purified RNC–channel complexes, J.-F.M. and C.W.A. obtained and analysed the electron microscopy data, J.C.G. helped with MDFF and channel modelling, S.J.L. helped with data analysis, W.L. and A.W. purified *M. jannaschii* components, and T.A.R., E.P. and C.W.A. wrote the paper.

Author Information Electron density maps have been submitted to the Electron Microscopy Data Bank (http://www.emdatabank.org/) under accession numbers EMD-5691, EMD-5692 and EMD-5693, and modelled structures to the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) under accession numbers 3J43, 3J44, 3J45, 3J46 and 1VVK. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.A.R. (Tom_Rapoport@hms.harvard.edu) or C.W.A. (cakey@bu.edu).

METHODS

Construction of plasmids and *E. coli* strains. Plasmids used in this study are listed and described in Supplementary Table 3. PCR reactions were performed with Phusion polymerase (New England Biolabs) or KOD polymerase (Novagen). The *E. coli* DH5α strain was used for all cloning procedures.

pBAD(MazF)-NC100, a plasmid expressing a SecM-stalled nascent chain under an arabinose-inducible (*ara*) promoter, and the MazF endoribonuclease under a tetracycline-inducible (*tet*) promoter, has been described[15]. In brief, a DNA sequence coding for a 100-amino-acid nascent chain was placed after the *ara* promoter of pBAD His/C (Invitrogen). The nascent chain contains an N-terminal signal sequence derived from *E. coli* DsbA, a Myc-tag and a C-terminal translational arrest sequence from *E. coli* SecM. The SecM nucleotide sequence contains three 'ACA' sites (underlined in the following sequence: 5′-TTCAGC<u>ACA</u>CCCGTCTGGATATC<u>ACA</u>AGC<u>ACA</u>AGGCATCCGTGCTGGCCCT-3′); MazF will cleave the mRNA at these positions and convert polysomes into monosomes. To keep MazF uninduced, a TetR repressor was expressed. The *tetR* gene from *Tn10* was cloned and inserted immediately downstream of the β-lactamase gene of the plasmid (for bicistronic expression). A DNA sequence for a *tet* promoter followed by *E. coli* MazF was cloned and placed between *tetR* and the replication origin of the plasmid. pACYC EhG/Y(68C), expressing a SecE–SecG fusion protein and SecY(68C) from a constitutive promoter, was constructed as follows. DNA sequences coding for *E. coli* SecE (residues 2–127) and SecG (residues 2–110) were fused with a sequence coding for a His-tag linker (GGSDGHHGHHHHGHHGDSGG). The fusion construct also contains an N-terminal calmodulin-binding peptide (CBP) tag (MGSRWKKNFIAVSAANRFKKISGGG). The resulting (CBP-tag)–SecE–(His-tag)–SecG fusion construct was ligated into pACYC-SecYEG[14], replacing the original SecE segment. Subsequently, the original SecG coding sequence from pACYC-SecYEG was removed by restriction enzyme digestion and re-ligation. For information on other plasmids, see Supplementary Table 3.

E. coli strains containing chromosome modifications were generated using standard λRed recombination techniques[31]. To construct an *E. coli* strain (EP71; BW25113 *Δrmf ΔompT rplL-strep::aadA(Str^R)*) in which ribosomal protein L12 (*rplL*) is C-terminally tagged with a *Strep*-tag (WSHPQFEK), we first synthesized a 'rplL-strep-RBS-aadA' DNA cassette, containing the C-terminal part of the *rplL* gene followed by the *Strep*-tag, a stop codon, a ribosome binding site (RBS), the coding sequence of a streptomycin resistance gene (*aadA*) and a short sequence downstream of the *rplL* gene. This cassette was amplified by PCR and electroporated into *ΔrmfΔompT* cells (EP51)[15] expressing λRed recombinase from the pKD46 plasmid. The resulting cells were selected on agar medium containing 25 μg ml^{-1} streptomycin. Incorporation of the cassette into the chromosome was verified by PCR and immunoblotting using *Strep*-tag antibodies (Novagen). To delete the chromosomal *secY* gene (strain EP72), EP71 cells were first transformed with pKD46 and pACYC EhG/Y(68C). After induction of λRed recombinase, the cells were electroporated with a PCR product containing a hygromycin resistance gene (*hph*), flanked by short sequences homologous to the chromosomal *secY* locus (so that the *secY* coding sequence is replaced by the *hph* coding sequence). Deletion of chromosomal *secY* was verified by PCR.

Preparation of SecY complex and ribosomes. All protein purification procedures were performed at 4 °C unless otherwise indicated. *M. jannaschii* and *E. coli* SecY complexes and *E. coli* 70S ribosomes were purified as described previously[3,8]. *M. jannaschii* cells were obtained from the University of Georgia Bioexpression and Fermentation Facility. *M. jannaschii* 70S ribosomes were purified by multiple ultracentrifugation steps as follows. Cells were homogenized in buffer containing 50 mM HEPES-NaOH, pH 7.5, 100 mM KCl, 10 mM $MgCl_2$ and 1 mM dithiothreitol (DTT), using a French press. After removing cell debris by centrifugation (SS34 rotor; 1 h at 16,000 r.p.m.), the cell homogenate was loaded onto a sucrose cushion (containing 50 mM HEPES, pH 7.5, 1 M NH_4Cl, 10 mM $MgCl_2$, 1 mM DTT and 30% (w/v) sucrose), and ribosomes were pelleted by ultracentrifugation at 45,000 r.p.m. for 5 h (Beckman Ti50.2 rotor). The pelleted ribosomes were re-suspended in buffer containing 50 mM HEPES, pH 7.5, 1 M NH_4Cl, 5 mM $MgCl_2$ and 1 mM DTT, and then sedimented by ultracentrifugation (SW-28 rotor, 24,000 r.p.m., 12 h) through a linear sucrose gradient (10–40% (w/v) sucrose in the re-suspension buffer). Fractions containing the 30S and 50S ribosomal subunits were collected separately and concentrated. The buffer was exchanged to 50 mM HEPES, pH 7.5, 100 mM NH_4Cl, 50 mM $MgCl_2$ and 1 mM DTT using a 100-kDa cut-off AmiconUltra (GE Healthcare) device. 30S and 50S subunits were mixed at a molar ratio of 2:1. To purify 70S ribosomes from excess 30S subunits, the complexes were subjected to centrifugation (SW-28 rotor, 24,000 r.p.m., 12 h) through a 10–40% sucrose gradient in 50 mM HEPES, pH 7.5, 100 mM NH_4Cl, 50 mM $MgCl_2$, 1 mM DTT. Fractions containing the 70S ribosomes were pooled, concentrated, and dialysed against buffer containing 50 mM HEPES, pH 7.5, 100 mM NH_4Cl, 10 mM $MgCl_2$ and 1 mM DTT. It should be noted that the resulting specimen contained an E-site tRNA at high occupancy.

Purification of disulphide-crosslinked *E. coli* RNC–SecY complexes. EP72 (*Δrmf ΔompT rplL-strep::aadA ΔsecY::hph* pACYC-EhG/Y(68C)) cells harbouring pBAD(MazF)-NC100 were grown to logarithmic phase in a medium containing 5 g l^{-1} trypton, 2.5 g l^{-1} yeast extract, 10 g l^{-1} casamino acids and 5 g l^{-1} NaCl. The expression of the nascent chain was induced by addition of 0.06% arabinose for 2 h at 37 °C, followed by *E. coli* MazF induction with 100 ng ml^{-1} anhydrotetracycline for 30 min at 30 °C. Disulphide crosslinking between NC100(19C) and SecY(68C) was then induced by addition of 1 mM 5,5′-dithiobis-(2-nitrobenzoic acid) (DTNB) to the culture medium for 20 min. DTNB facilitates disulphide-bond formation between SecY and the nascent chain as efficiently as Cu-phenanthroline ($CuPh_3$)[15]. The cells were pelleted, washed once with buffer containing 50 mM Tris-HCl, pH 7.2, 5 mM $Mg(OAc)_2$, 150 mM KCl, and frozen. RNC–SecY complexes were purified as follows. The cells were re-suspended in buffer containing 50 mM Tris-acetate, pH 7.2, 25 mM $Mg(OAc)_2$, 0.3 M NH_4Cl and homogenized with a French press. One per cent *n*-dodecyl-β-D-maltoside (DDM) was added to the cell lysate for 1 h to solubilize membranes. After centrifugation (SS-34 rotor, 13,000 r.p.m., 30 min), ribosomes containing *Strep*-tagged L12 were purified by applying the lysate to a Strep-Tactin Sepharose column (IBA). The column was washed with 8 column volumes (CV) of buffer containing 50 mM Tris-acetate, pH 7.2, 25 mM $Mg(OAc)_2$, 0.4 M NH_4Cl, 0.03% DDM, and then with 2 CV of buffer (TMP200) containing 50 mM Tris-acetate, pH 7.2, 25 mM $Mg(OAc)_2$, 0.2 M KOAc, 0.03% DDM. Ribosomes were eluted from the column with 4 CV of the TMP200 buffer containing 4 mM desthiobiotin. To enrich for channel-bound RNCs containing His-tagged SecE–SecG fusion protein, the eluate was incubated with Dynal-Talon beads (Invitrogen) for 30 min. The beads were washed three times with TMP200 buffer, and bound complexes were eluted with TMP200 buffer containing 120 mM imidazole. The complexes were further purified by gel filtration on a Superose 6 column (GE Healthcare) equilibrated with buffer containing 50 mM Tris-acetate, pH 7.2, 10 mM $Mg(OAc)_2$, 80 mM KOAc, 0.03% DDM. Monomeric ribosome fractions were collected and concentrated to 8–9 mg ml^{-1}.

Test for *in vitro* reconstitution of the RNC–SecY complex. For the experiments shown in Supplementary Fig. 12, RNCs containing the $DsbA108_{His}$ or NC100 nascent chain were isolated as follows. pBAD-$DsbA108_{His}$(19C) or pBAD-NC100(19C) were transformed into *ΔrmfΔompT* cells (EP51) harbouring the pRARE2 plasmid. Cells were grown to log phase in 2 × YT medium (16 g l^{-1} tryptone, 10 g l^{-1} yeast extract and 5 g l^{-1} NaCl) supplemented with 100 μg ml^{-1} ampicillin and 40 μg ml^{-1} chloramphenicol. Nascent chain expression was induced by addition of 0.4% arabinose for 3 h. The cells were re-suspended in buffer (TMA750) containing 50 mM Tris-acetate, pH 7.2, 25 mM $Mg(OAc)_2$, 0.75 M NH_4Cl and 1.5 mM DTT and homogenized in a French press. To solubilize the membranes, 1% DDM was added to the cell extract. The extract was cleared by centrifugation at 13,000 r.p.m. for 1 h. The ribosomes were sedimented through a sucrose cushion (TMA750, 30% sucrose, 0.03% DDM) and re-suspended in TMA750. The buffer was exchanged on a PD-10 desalting column (GE Healthcare) to buffer TMP100 (50 mM Tris-acetate, pH 7.2, 25 mM $Mg(OAc)_2$, 0.1 M KOAc). To purify RNCs containing monosomes, the ribosomes (OD_{260nm} = 500–1,000) in TMP750 were briefly incubated with 20 μg ml^{-1} RNase A at room temperature (23 °C) and immediately injected into a Superose 6 gel-filtration column (GE Healthcare) equilibrated with TMP100 containing 50 mM Tris-acetate, pH 7.2, 25 mM $Mg(OAc)_2$ and 100 mM KOAc. Fractions containing monomeric ribosomes were collected.

$DsbA108_{His}$- or NC100-containing RNCs (0.27 μM total ribosomes) were mixed with a 15-fold excess (4.1 μM) of the SecY(68C) complex in TMP100 containing 0.03% DDM. When SecY–nanodiscs were used instead of SecY–detergent complexes, 0.138 μM of RNCs were mixed with a fivefold (0.7 μM) excess of SecY–nanodiscs in the same buffer lacking detergent. After incubating solutions at 4 °C for 1 h or at 30 °C for 30 min, disulphide bridge formation was induced by addition of 0.1 mM $CuPh_3$ for 20 min at room temperature. The reaction was stopped by addition of 20 mM *N*-ethyl maleimide for 30 min at 4 °C. The samples were subjected to non-reducing SDS–PAGE and analysed by immunoblotting with Myc and SecY antibodies.

Nanodiscs containing SecY(68C) complex were generated as previously described[32] using the scaffold protein MSP1D1 (ref. 33). In brief, SecY(68C) complexes, MSP1D1 and deoxyBigChap-solubilized *E. coli* polar lipid (Avanti Polar Lipids) were mixed in a molar ratio of 1:4:100 in 50 mM Tris-acetate, pH 7.2, 150 mM KOAc. After removal of the detergent with Biobeads (Bio-Rad), the sample was injected into a Superdex 200 column equilibrated with buffer TMP100. Fractions containing the SecY–nanodiscs complex were pooled and concentrated with an Amicon Ultra device (100-kDa cut-off).

SDS–PAGE and immunoblotting. SDS–PAGE was performed using 4–12% Bis-Tris gels (Bio-Rad) with either MES-SDS or MOPS-SDS running buffer (Invitrogen). Images of immunoblots were recorded with a charge-coupled device (CCD)-based device (Fujifilm LAS-3000) and a standard ECL reagent. Antibodies against the

C terminus of SecY were described previously[34]. Anti-Myc and anti-CBP antibodies were obtained from Sigma and Genscript, respectively.

Cryo-electron microscopy and three-dimensional image processing. *M. jannaschii* ribosomes were mixed with a fivefold excess of *M. jannaschii* SecYEβ in 100 mM NH_4Cl, 30 mM $MgCl_2$, 20 mM HEPES-KOH, pH 7.5, 6 mM β-mercaptoethanol and ~0.1% DDM. Samples were added to 400 mesh Cu grids with a holey carbon film (Quantafoil 2/1); ~2 μl per grid at an OD_{260} of 60–120 or diluted and added to 400 mesh grids with a thin continuous carbon film. After blotting, samples were plunge frozen into liquid ethane with a Vitrobot Mark 3 (FEI). Grids were mounted on an Oxford cold holder and imaged at 200 kV on a Tecnai F20. Data were collected manually on Kodak SO163 film at ×50,000 with a defocus range of −1.0 to −2.5 μm. Micrographs were scanned on Zeiss SCAI and Creoscitex EVERSMART scanners and particles selected with EMAN boxer[35] were binned and scaled to 2.73 Å per pixel. In total, ~59,000 particles were corrected for the contrast transfer function (CTF) with EMAN2 and classified with a supervised multi-reference refinement into groups, with and without channel, to give a data set with ~37,000 particles that contained the channel. Three-dimensional reconstructions from six EMAN2 refinements carried out with different parameters and estimated resolutions of 9.2–9.5 Å (based on half-data set comparisons) were aligned in Chimera and averaged to obtain a final three-dimensional density map.

Non-programmed *E. coli* ribosome–channel complexes were prepared for cryo-electron microscopy and imaged at ×50,000 with a Gatan (626-DH) cold holder at 200 kV, as described previously[8]. After identifying and removing complexes without channels, ~39,000 particles were processed with EMAN1 (ref. 35) at a pixel size of 2.73 Å (for details see ref. 8). Aliquots of *E. coli* RNCs with SecYEG (OD_{260} = 120–160 in ~0.06–0.1% DDM) were thawed and kept on ice. Samples were applied to 300 mesh Cu grids with a holey support film (Quantafoil 2/1 for imaging at ×42,000) and 400 mesh grids (Quantafoil 1.2/1.3 for imaging at ×50,000). The holey grids had a very thin layer of carbon freshly applied by evaporation and were airglow discharged before use. A Vitrobot or a manual plunger was used to plunge-freeze grids after blotting into liquid ethane, with the chamber at room temperature and a relative humidity of ~95–100%. Samples were loaded onto an Oxford cold holder and images obtained at 160 kV on a 4096 × 4096 CCD (TVIPS) with a semi-automated, single-particle collection program in EMtools (TVIPS) on a TF-20. Particle images were selected using e2boxer and further processed with EMAN2 (ref. 29).

The CTF correction was based on all particles from each CCD frame (~450,000 from ~3500 frames), including RNC–channel complexes that formed aggregates, after scaling data collected at ×50,000 to 2.12 Å per pixel. Subsequently, multiple cycles of reference free classification in EMAN2 were used to extract ~167,000 single particles without close nearest neighbours for final processing. A ribosome at 25 Å resolution, with and without the channel, was used as a starting model. The program e2refinemulti.py was used to separate the data set into two groups, which were refined separately to a resolution of ~11–12 Å. A final supervised classification with e2refinemulti.py at an angular step size appropriate for 14 Å resolution was then carried out with the full data set, using three-dimensional references with and without the channel filtered to 14 Å. This step used the Fourier ring correlation comparator and provided an improved separation of the data set. At this stage ~83,000 particles with channels from the supervised classification were sorted further with e2ligandclassify.py, on the basis of their signal-to-noise ratio, to give a final data set of ~53,000 particles. Two separate structure refinements were then done, starting with either the best three-dimensional reference from the original low-resolution ribosome model or using a 6.8 Å resolution *E. coli* ribosome map (EMDB code, 5036) scaled to 2.12 Å per pixel. After convergence, the four best maps (two from each structure path calculated with different refinement parameters) were aligned in Chimera and averaged to give the final three-dimensional map.

Molecular modelling and docking. Maps from *M. jannaschii* and active *E. coli* ribosome–channel complexes were subjected to a local normalization in EMAN2 to allow densities for ribosomal proteins, RNA, channel and micelle to be displayed and analysed using a single density cut-off. Maps were segmented with Chimera using *Zone* and difference map options (*vop subtract*)[30]. Small and large ribosomal subunit models were fit into the ribosome–channel density maps using Chimera *fit in map* option[30] and MDFF[7] with runs of 500,000 steps (0.5 ns). Because no model was available for the *M. jannaschii* ribosome, we used a model of the related complex from *P. furiosus* (ref. 6, PDB ID, 3J20, 3J21 and 3J2L). Extra copies of ribosomal proteins and ribosomal RNA loops from the *P. furiosus* model that are absent in *M. jannaschii* were omitted. For *E. coli* ribosome–channel complexes, a nearly complete model of the large ribosomal subunit based on electron microscopy modelling and a crystal structure (ref. 11, PDB ID: 3J01; ref. 12, PDB ID: 2I2T) were used, along with a crystal structure of the small subunit (ref. 12, PDB ID: 2I2P). Models for tRNAs and mRNA were obtained from a crystal structure of a programmed *T. thermophilus* ribosome (ref. 36, PDB ID: 3I8G).

The global resolution in experimental density maps was determined separately for the ribosome and channel in each structure using Fourier shell correlation (FSC) in EMAN2, with reference maps calculated from Protein Data Bank files of docked models. Reference maps were calculated with pdb2mrc in EMAN at 7 Å resolution and aligned in Chimera to the appropriate experimental map, then saved with *vop resample onGrid*. Experimental maps of ribosomes, as part of their cognate ribosome–channel complex, had a soft mask applied after calculation in EMAN2. Density maps for channels were created by segmentation in Chimera which also effectively created a mask. However, no masks were created for reference maps to prevent spurious correlations between similar masks in the FSC calculations between the two volumes being compared. The 0.5 criterion was used in all cases to identify the resolution.

Models for closed and open *E. coli* SecYEG channels were constructed as follows. SecY in the closed channel was based on individual structural elements (helices and turns) from the crystal structure of *T. thermophilus* SecY[4]. These segments were docked onto the closed crystal structure of SecY from *M. jannaschii*[3] in Chimera, on the basis of sequence alignments between the three organisms. Loops were then regularized and additional residues added as needed in Coot[37]. SecE and SecG subunits were taken from the crystal structure of *T. maritima* SecYEG[13]. The structural model was then mutated to *E. coli* sequences, energy minimized with NAMD[38] and fit into the map with Chimera[30] and MDFF[7]. A model for the open *E. coli* channel was constructed in a similar way, on the basis of a crystal structure of a partially open SecYE channel from *P. furiosus*[5]. SecY models were positioned initially in the maps by docking the 6/7 and 8/9 loops into their density with Rosetta[39]. All MDFF runs with these components were done with segmented maps that contained the large ribosomal subunit and complete density for the channel and micelle. Models for the large subunit and SecY channel were minimized together. Importantly, the model of a partially open channel moved into correct density, to reveal the signal sequence helix and associated density for the nascent chain. Finally, no density was observed for the first two TMs of *E. coli* SecE, which are connected by an extended linker to the surface helix of this subunit, and thus may be flexible.

31. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97,** 6640–6645 (2000).
32. Alami, M., Dalal, K., Lelj-Garolla, B., Sligar, S. G. & Duong, F. Nanodiscs unravel the interaction between the SecYEG channel and its cytosolic partner SecA. *EMBO J.* **26,** 1995–2004 (2007).
33. Nath, A., Atkins, W. M. & Sligar, S. G. Applications of phospholipid bilayer nanodiscs in the study of membranes and membrane proteins. *Biochemistry* **46,** 2059–2069 (2007).
34. Cannon, K. S., Or, E., Clemons, W. M. Jr, Shibata, Y. & Rapoport, T. A. Disulfide bridge formation between SecY and a translocating polypeptide localizes the translocation pore to the center of SecY. *J. Cell Biol.* **169,** 219–225 (2005).
35. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128,** 82–97 (1999).
36. Jenner, L. B., Demeshkina, N., Yusupova, G. & Yusupov, M. Structural aspects of messenger RNA reading frame maintenance by the ribosome. *Nature Struct. Mol. Biol.* **17,** 555–560 (2010).
37. Emsley, P. & Cowtan, K. *Coot*: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
38. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).
39. DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W. & Baker, D. Refinement of protein structures into low-resolution density maps using Rosetta. *J. Mol. Biol.* **392,** 181–190 (2009).

LETTER

doi:10.1038/nature12950

Structures of the Sec61 complex engaged in nascent peptide translocation or membrane insertion

Marko Gogala[1], Thomas Becker[1], Birgitta Beatrix[1], Jean-Paul Armache[1], Clara Barrio-Garcia[1], Otto Berninghausen[1] & Roland Beckmann[1]

The biogenesis of secretory as well as transmembrane proteins requires the activity of the universally conserved protein-conducting channel (PCC), the Sec61 complex (SecY complex in bacteria)[1]. In eukaryotic cells the PCC is located in the membrane of the endoplasmic reticulum where it can bind to translating ribosomes for co-translational protein transport. The Sec complex consists of three subunits (Sec61α, β and γ) and provides an aqueous environment for the translocation of hydrophilic peptides as well as a lateral opening in the Sec61α subunit that has been proposed to act as a gate for the membrane partitioning of hydrophobic domains[2]. A plug helix and a so-called pore ring are believed to seal the PCC against ion flow and are proposed to rearrange for accommodation of translocating peptides[2,3]. Several crystal and cryo-electron microscopy structures revealed different conformations of closed and partially open Sec61 and SecY complexes[2,4–8]. However, in none of these samples has the translocation state been unambiguously defined biochemically. Here we present cryo-electron microscopy structures of ribosome-bound Sec61 complexes engaged in translocation or membrane insertion of nascent peptides. Our data show that a hydrophilic peptide can translocate through the Sec complex with an essentially closed lateral gate and an only slightly rearranged central channel. Membrane insertion of a hydrophobic domain seems to occur with the Sec complex opening the proposed lateral gate while rearranging the plug to maintain an ion permeability barrier. Taken together, we provide a structural model for the basic activities of the Sec61 complex as a protein-conducting channel.

Ribosome–protein-conducting channel (PCC) complexes were formed using a well-established *in vitro* protein translocation system: a wheat germ translation extract combined with canine pancreatic endoplasmic reticulum membranes[9]. We chose nascent polypeptides derived from the leader peptidase (Lep) protein as intermediates. Dependent on the hydrophobicity of a variable region they show distinct translocation (LepT) or membrane insertion (LepM) behaviour[10] (Fig. 1a). The peptides are 338 amino acids long, carry two transmembrane helices (TM1, TM2) followed by a 149 amino acid long lumenal loop containing a first glycosylation (GS1) site as well as a streptavidin and a haemagglutinin tag. The loop ends with the variable region that is either hydrophilic (LepT) or hydrophobic (LepM), followed by a carboxy-terminal stretch containing the second glycosylation site (GS2) and a ribosomal stalling sequence (cytomegalovirus (CMV) upstream open reading frame (uORF) *gp48*) (Fig. 1a)[11,12]. The length of the C-terminal stretch (93 amino acids) was chosen such that the variable region can fully engage the PCC when translation is stalled. Notably, the translocation state of these peptides can be precisely monitored by their glycosylation state: only one of two possible sites, that on the lumenal side (GS1) of the endoplasmic reticulum membrane, is glycosylated when the nascent peptide is trapped in the ribosome–PCC complex in an intermediate state. Hence, this novel approach should provide us with bona fide translocation intermediates that are functionally better defined than complexes that have been reconstituted *in vitro* from purified complexes.

When the messenger RNAs were translated in the absence of membranes, we observed the expected stalled transfer-RNA-bound peptide as well as free peptide owing to ineffective stalling (Fig. 1b, c). In the presence of membranes, however, after translocation of the loop region a single glycosylation event led to one additional shifted peptidyl-tRNA band for both constructs, LepT and LepM. Because of inefficient stalling two additional bands indicated dual and single glycosylation events, respectively, for the free fully translocated LepT and membrane-inserted LepM peptides (Fig. 1b, c). The peptidyl-tRNA bands disappear after puromycin treatment as expected (Fig. 1b). After optimizing the translation conditions for enrichment of stalled and glycosylated intermediates, ribosome- and membrane-bound nascent peptides were isolated by membrane pelleting and mild detergent solubilisation followed by affinity purification via the streptavidin-tag in the nascent peptide. Western blot analysis of the purified sample indicated that the final fractions consisted of ribosome–nascent chain complexes (RNCs) highly enriched for mono-glycosylated tRNA-bound nascent peptide and Sec61 complex (Fig. 1b, c and Extended Data Fig. 1), which was confirmed by mass spectrometry. The purified complexes therefore represented bona fide translocation or membrane insertion intermediates.

The purified LepT and LepM complexes were subjected to cryo-electron microscopy and single-particle analysis. Applying *in silico* sorting procedures to the data sets resulted in final reconstructions of the two programmed RNC–Sec61 complexes as well as an idle ribosome–Sec61 complex lacking peptidyl-tRNA. All structures were solved at resolutions between 6.9 and 7.8 Å (Fig. 1d-f and Extended Data Fig. 2). Resolution measurements were further validated by measuring cross-resolution between channel densities and the obtained molecular models (Extended Data Fig. 3), resulting in a local resolution of approximately 7.5 Å for all Sec61 densities, which allowed unambiguous resolution of α-helical secondary structure in the channels.

The presence of peptidyl-tRNA indicated a high degree of programming for the LepT and LepM intermediates. The densities corresponding to the PCC showed the central Sec61 protein surrounded by a mixed detergent lipid micelle essentially as observed before[7]. The transmembrane segments, the proposed lateral gate and the plug helix of Sec61α were well resolved and, thus, allowed for unambiguous positioning of homology models in all cases (Extended Data Fig. 4). Notably, an extra density belonging to an inserting transmembrane helix of LepM was observed in the lateral gate of the LepM-engaged PCC (Extended Data Fig. 4c).

Overall, the mode of ribosome binding seemed to be very similar in all three structures (Extended Data Fig. 4). Cytoplasmic loops L6/L7 and L8/L9 of Sec61α contact the universal ribosomal adaptor site, consistent with previous cryo-electron microscopy studies of ribosome-bound Sec61 and SecY complexes[7,13,14]. This indicates that, in eukaryotes, the overall mode of ribosome binding is not directly dependent on the different modes of PCC activity.

The conformation of the ribosome-bound idle Sec61 complex bears a strong resemblance to the conformation observed in the archaeal *Methanococcus jannaschii* SecYEβ crystal structure[2], in which the

[1]Gene Center and Center for integrated Protein Science Munich, Department of Biochemistry, Feodor-Lynen-Strasse 25, University of Munich, 81377 Munich, Germany.

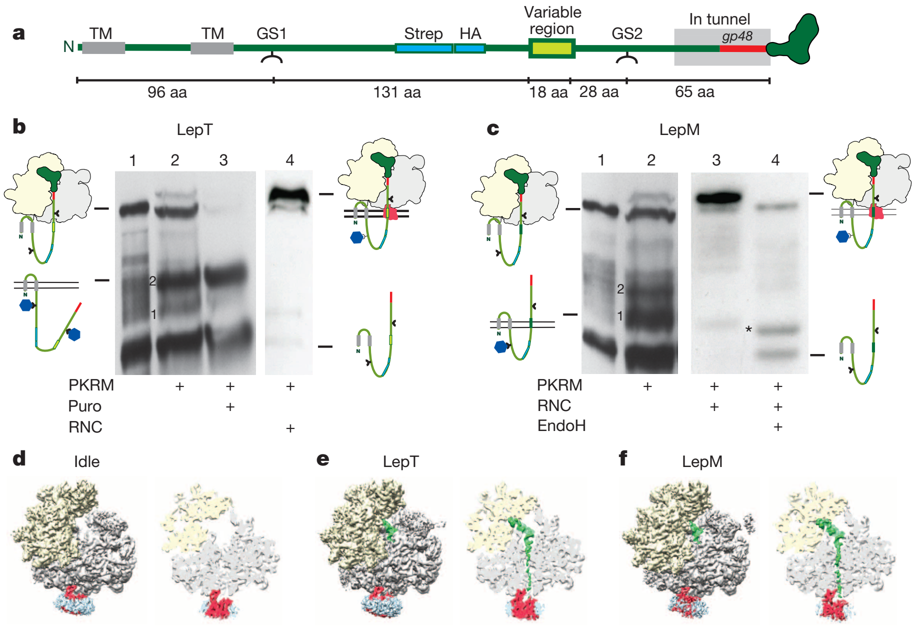


Figure 1 | Generation and cryo-electron microscopy structures of translocating and inserting ribosome–Sec61 complexes. **a**, Line diagram of LepT and LepM constructs with distinct variable regions. aa, amino acids; HA, haemagglutinin; Strep, streptavidin. **b**, **c**, Analysis and purification of LepT and LepM intermediates. Western blots probing for haemagglutinin indicate (glycosylated) peptidyl-tRNA, unglycosylated as well as mono- and bi-glycosylated free peptides as illustrated in schematic drawings. An EndoH background band is indicated by an asterisk. **d**–**f**, Cryo-electron microscopy reconstructions of the idle 80S–Sec61 complex (**d**), the LepT–RNC–Sec61 complex (**e**) and the LepM–RNC–Sec61 complex (**f**). Right panels show cut density to visualize the ribosomal tunnel and peptidyl-tRNA.

lateral gate is closed and the plug obstructs the central constriction of the channel. In the idle Sec61 complex the lateral gate is also closed (Fig. 2a); however, with the lumenal part of TM7 shifted slightly towards the amino-terminal half of Sec61 and with the plug also shifted by approximately 3.5 Å towards the lumenal side when compared to the crystal structure. This movement is accompanied by a small rotation of TM10 of Sec61α (Extended Data Fig. 5a, b) and may explain changes in ion conductivity observed upon ribosome binding to Sec61[15].

In the engaged Sec61 complex containing the hydrophilic LepT intermediate, the conformation of Sec61α is slightly more open compared to the idle state, and is very similar to a previous cryo-electron microscopy structure of a RNC-bound mammalian Sec61 complex[7] (Extended Data Fig. 5f). The lateral gate is partially opened, yet only by a 4 Å lateral shift of TM7 of Sec61α (Fig. 2c). The observed rather small opening of the two halves of Sec61α during peptide translocation is in agreement with chemical crosslink studies[16]. Here, TM7 and TM2 of a translocating

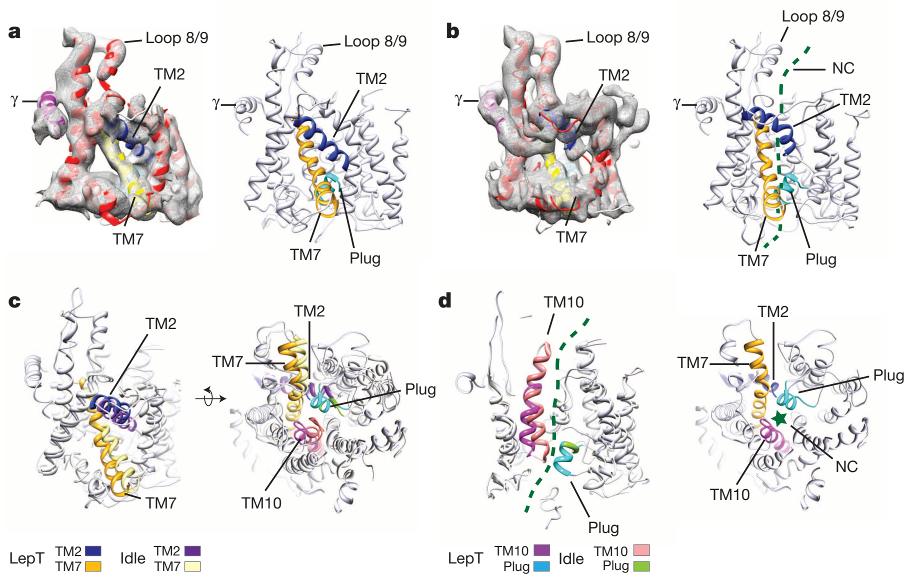


Figure 2 | Models for idle and translocating Sec61. **a**, Model for the idle Sec61 complex with (left) and without isolated density (right). **b**, Model for the translocating LepT-engaged Sec61 complex. NC, nascent chain. **c**, **d**, Comparison between idle and LepT-engaged Sec61 complex. **c**, Left, view on the lateral gate; right, lumenal view focusing on the plug; **d**, left, side view focusing on TM10 and the plug; right, cytoplasmic view of the LepT-engaged Sec61 complex. For LepT models the presence of the translocating peptide is indicated as a green dashed line or a green asterisk. The colour code for TM2, TM7, TM10 and the plug is given underneath.

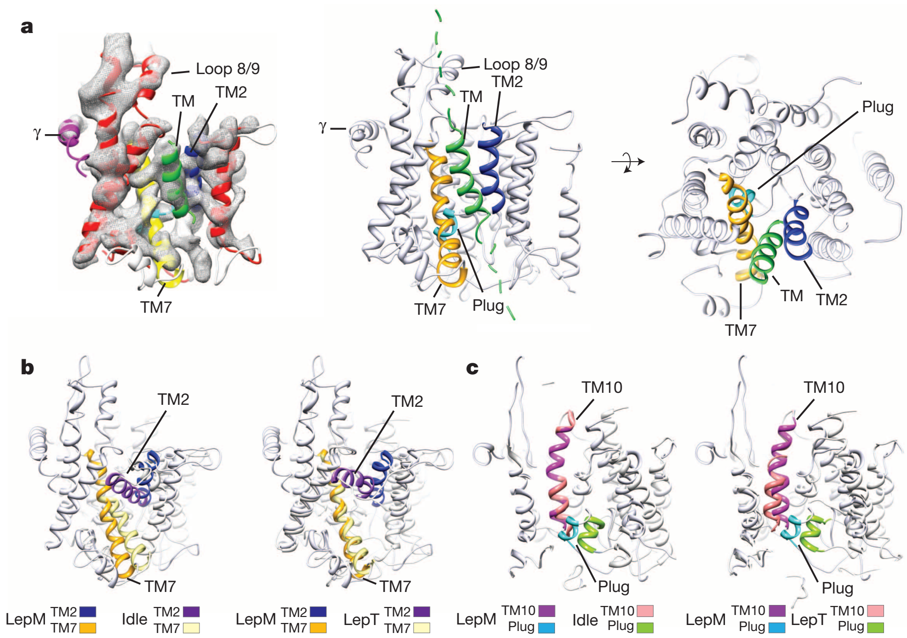


Figure 3 | Model for the membrane inserting Sec61. **a**, Model for the inserting LepM-engaged Sec61 complex with isolated density (left); middle and right, side view and cytoplasmic view of the Sec61 complex. **b**, **c**, Comparison between LepM-engaged and idle Sec61 complex (left) and LepT-engaged Sec61 complex (right) focusing on the lateral gate (**b**) and focusing on TM10 and the plug (**c**). The model for the inserting LepM transmembrane helix (TM) is shown in green in **a**. The colour code for TM2, TM7, TM10 and the plug is given underneath.

SecY can be locked using cross-linkers with very short spacer length (between 2 and 5 Å), strongly indicating that a rather closed conformation of the lateral gate still allows for peptide translocation[16].

In our LepT complex the plug did not show a detectable shift when compared to the idle Sec61 complex. However, immobilization of the plug has been shown to still allow unrestricted protein translocation in bacteria[17]. In the LepT-engaged Sec61α the lumenal part of TM10 was shifted outward, away from the plug (Fig. 2d). This shift of approximately 6 Å would be sufficient to provide the required opening for the accommodation of an extended translocating peptide segment between the plug and TM10[18]. This is in perfect agreement with previous cross-link data showing that the translocating peptide is in the immediate vicinity of the plug helix, TM10 and TM5 of the SecY complex. Moreover, this shift would also disrupt the aliphatic pore-ring by pulling one participating residue from TM10 out of the assembly. At the given resolution the extended and most likely flexible translocating peptide was not visible. Regardless, the observed conformation allows placing an arbitrary model peptide, even with bulky side chains, into the aqueous interior of the PCC that traverses from the cytoplasmic to the lumenal side without any clashes (Extended Data Fig. 6). Furthermore we believe that subtle changes in the position of helices in the central channel (for example, TM10) can dynamically accommodate the geometry of virtually any translocating peptide.

Taken together, when engaged in translocation of a hydrophilic nascent peptide, the Sec61 complex can adopt a conformation with a lateral gate opened only by a few Ångstroms and a continuous central conduit provided without major displacement of the plug.

In the Sec61 complex containing the hydrophobic LepM intermediate the conformation of the complex showed an open lateral gate (Fig. 3, Extended Data Fig. 4c). TM2 and TM7 of Sec61α moved apart by approximately 12 Å to create a gap that harbours a rod-like extra density corresponding to three to four turns of an α-helix. Previous biochemical data[16,19] and cryo-electron microscopy studies of the SecYE complex[8,20] (Extended Data Fig. 5g, h) indicated the position of an engaged signal sequences or a signal anchor sequence in the lateral gate. Furthermore, it has been shown that canonical transmembrane domains can also be retained at the PCC by protein–protein interactions until release is triggered by translation termination or the arrival of another transmembrane segment[21–24]. It is therefore likely that the observed density represents the helical LepM transmembrane segment that is still connected by the C-terminal linker to the tRNA. This would support the hypothesis that nascent transmembrane domains are indeed inserted into the lipid bilayer through the proposed lateral gate[2,19,25] and that these domains adopt a helical conformation when partitioning from the PCC into the lipid phase[26–28].

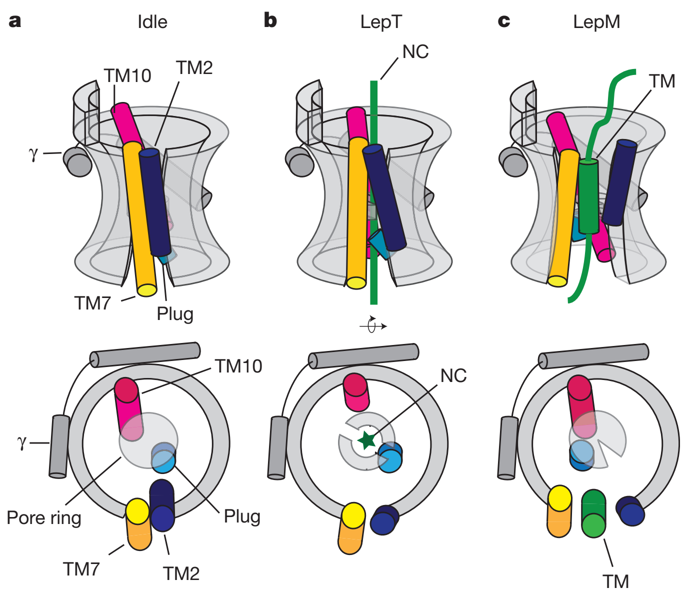


Figure 4 | Conformational transitions of Sec61 during co-translational protein translocation and membrane insertion. **a**, In the ribosome-bound idle state, the lateral gate of the Sec61 complex is closed and the central constriction is closed by TM10 and the plug. **b**, When engaged with a translocating peptide the lumenal part of TM10 moves outward. This creates a central opening for the hydrophilic nascent chain (green). Whereas the plug and TM7 remain unchanged, TM2 rearranges slightly, resulting in a partial opening of the lateral gate. **c**, Upon encounter of a more hydrophobic peptide stretch that is supposed to be inserted into the lipid bilayer as a transmembrane domain (stop-transfer sequence), the lateral gate opens up further. It accommodates the peptide segment in a helical conformation between TM2 and TM7 to allow access to the lipid phase. Lateral gate opening and transfer of the hydrophobic peptide from the central aqueous channel into the lateral gate are accompanied by a concomitant inwards movement of the plug and TM10. Thereby, the ion permeability barrier may be maintained.

In the observed open conformation of the Sec61 complex the plug has moved towards the central constriction of the channel (Figs 3c, 4). This is consistent with fluorescence data indicating that the plug does not move into a more hydrophilic environment during transmembrane helix insertion[29]. In its new position the plug closes the void created by both the exiting peptide and the separation of the N- and C-terminal halves of Sec61α. In addition, an inward movement of TM10 towards the plug by 3 Å was observed (Fig. 3c). This concomitant movement of plug and TM10 may contribute to maintain a sealed central channel when a hydrophobic stretch arrives at the PCC and leaves the aqueous interior to engage the lateral gate (Fig. 4). Notably, the conformation of the open Sec61 complex is most similar to that observed in the crystal structure of an idle archaeal Sec complex[6] (Extended Data Fig. 5i) in which the crystal packing led to an opening of the lateral gate.

Taken together, the specifically glycosylated stalled nascent peptides provide a solid biochemical foundation for our structural analysis of engaged ribosome-bound protein-conducting channels. Our models provide a basic structural framework on the conformational transitions that enable the Sec61α to function in peptide translocation as well as in membrane insertion of nascent polypeptides (Fig. 4).

METHODS SUMMARY

Bona fide translocating or inserting RNC–Sec61 complexes containing glycosylated LepT and LepM nascent peptides were generated using a wheat germ cell-free extract supplemented with dog pancreas membranes (puromycin/high-salt-treated rough membranes, PKRM) and purified signal recognition particle (SRP). Ribosome-containing membranes were pelleted, solubilized with digitonin and RNC–PCC complexes were affinity-purified essentially as described before[30] using a streptavidin-tag on the nascent peptide (see Methods). For cryo-electron microscopy, LepT–RNC–Sec61 and LepM–RNC–Sec61 complexes were vitrified and data were collected on a Titan Krios electron microscope (FEI). Single-particle analysis, three dimensional reconstruction and computational sorting were done using the SPIDER software package[31]. Molecular modelling and visualization was done using the Coot[32] and Chimera[33] software packages.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 June; accepted 10 December 2013.

1. Park, E. & Rapoport, T. A. Mechanisms of Sec61/SecY-mediated protein translocation across membranes. *Annu. Rev. Biophys.* **41,** 21–40 (2012).
2. van den Berg, B. *et al.* X-ray structure of a protein-conducting channel. *Nature* **427,** 36–44 (2004).
3. Park, E. & Rapoport, T. A. Preserving the membrane barrier for small molecules during bacterial protein translocation. *Nature* **473,** 239–242 (2011).
4. Zimmer, J., Nam, Y. & Rapoport, T. A. Structure of a complex of the ATPase SecA and the protein-translocation channel. *Nature* **455,** 936–943 (2008).
5. Tsukazaki, T. *et al.* Conformational transition of Sec machinery inferred from bacterial SecYE structures. *Nature* **455,** 988–991 (2008).
6. Egea, P. F. & Stroud, R. M. Lateral opening of a translocon upon entry of protein suggests the mechanism of insertion into membranes. *Proc. Natl Acad. Sci. USA* **107,** 17182–17187 (2010).
7. Becker, T. *et al.* Structure of monomeric yeast and mammalian Sec61 complexes interacting with the translating ribosome. *Science* **326,** 1369–1373 (2009).
8. Frauenfeld, J. *et al.* Cryo-EM structure of the ribosome–SecYE complex in the membrane environment. *Nature Struct. Mol. Biol.* **18,** 614–621 (2011).
9. Erickson, A. H. & Blobel, G. Cell-free translation of messenger RNA in a wheat germ system. *Methods Enzymol.* **96,** 38–50 (1983).
10. Sääf, A., Wallin, E. & von Heijne, G. Stop-transfer function of pseudo-random amino acid segments during translocation across prokaryotic and eukaryotic membranes. *Eur. J. Biochem.* **251,** 821–829 (1998).
11. Bhushan, S. *et al.* Structural basis for translational stalling by human cytomegalovirus and fungal arginine attenuator peptide. *Mol. Cell* **40,** 138–146 (2010).
12. Degnin, C. R., Schleiss, M. R., Cao, J. & Geballe, A. P. Translational inhibition mediated by a short upstream open reading frame in the human cytomegalovirus gpUL4 (gp48) transcript. *J. Virol.* **67,** 5514–5521 (1993).
13. Ménétret, J.-F. *et al.* Single copies of Sec61 and TRAP associate with a nontranslating mammalian ribosome. *Structure* **16,** 1126–1137 (2008).
14. Ménétret, J.-F. *et al.* Ribosome binding of a single copy of the SecY complex: implications for protein translocation. *Mol. Cell* **28,** 1083–1092 (2007).
15. Simon, S. M. & Blobel, G. A protein-conducting channel in the endoplasmic reticulum. *Cell* **65,** 371–380 (1991).
16. du Plessis, D. J., Berrelkamp, G., Nouwen, N. & Driessen, A. J. The lateral gate of SecYEG opens during protein translocation. *J. Biol. Chem.* **284,** 15805–15814 (2009).
17. Lycklama a Nijeholt, J. A., Bulacu, M., Marrink, S. J. & Driessen, A. J. Immobilization of the plug domain inside the SecY channel allows unrestricted protein translocation. *J. Biol. Chem.* **285,** 23747–23754 (2010).
18. Cannon, K. S., Or, E., Clemons, W. M. Jr, Shibata, Y. & Rapoport, T. A. Disulfide bridge formation between SecY and a translocating polypeptide localizes the translocation pore to the center of SecY. *J. Cell Biol.* **169,** 219–225 (2005).
19. Plath, K., Mothes, W., Wilkinson, B. M., Stirling, C. J. & Rapoport, T. A. Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell* **94,** 795–807 (1998).
20. Hizlan, D. *et al.* Structure of the SecY complex unlocked by a preprotein mimic. *Cell Rep.* **1,** 21–28 (2012).
21. Sadlish, H., Pitonzo, D., Johnson, A. E. & Skach, W. R. Sequential triage of transmembrane segments by Sec61α during biogenesis of a native multispanning membrane protein. *Nature Struct. Mol. Biol.* **12,** 870–878 (2005).
22. Pitonzo, D., Yang, Z., Matsumura, Y., Johnson, A. E. & Skach, W. R. Sequence-specific retention and regulated integration of a nascent membrane protein by the endoplasmic reticulum Sec61 translocon. *Mol. Biol. Cell* **20,** 685–698 (2009).
23. Hou, B., Lin, P. J. & Johnson, A. E. Membrane protein TM segments are retained at the translocon during integration until the nascent chain cues FRET-detected release into bulk lipid. *Mol. Cell* **48,** 398–408 (2012).
24. Heinrich, S. U. & Rapoport, T. A. Cooperation of transmembrane segments during the integration of a double-spanning protein into the ER membrane. *EMBO J.* **22,** 3654–3663 (2003).
25. Martoglio, B., Hofmann, M. W., Brunner, J. & Dobberstein, B. The protein-conducting channel in the membrane of the endoplasmic reticulum is open laterally toward the lipid bilayer. *Cell* **81,** 207–214 (1995).
26. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433,** 377–381 (2005).
27. Hessa, T. *et al.* Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **450,** 1026–1030 (2007).
28. Heinrich, S. U., Mothes, W., Brunner, J. & Rapoport, T. A. The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell* **102,** 233–244 (2000).
29. Lycklama a Nijeholt, J. A., Wu, Z. C. & Driessen, A. J. Conformational dynamics of the plug domain of the SecYEG protein-conducting channel. *J. Biol. Chem.* **286,** 43881–43890 (2011).
30. Halic, M. *et al.* Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature* **427,** 808–814 (2004).
31. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116,** 190–199 (1996).
32. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
33. Pettersen, E. F. *et al.* UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

Acknowledgements We thank C. Ungewickell for assistance with cryo-electron microscopy, E. van der Sluis for discussions, S. Funes for help with reagent preparation, B. Dobberstein for endoplasmic reticulum membranes, F. Förster and S. Pfeffer for performing mass spectrometry analysis. This work was supported by grants of the German Research Council (SFB594 to R.B. and B.B., SFB646 to T.B. and R.B., GRK1721 to R.B.). R.B. acknowledges support by the Center for Integrated Protein Science and the European Research Council (Advanced Grant CRYOTRANSLATION).

Author Contributions M.G., T.B. and R.B. designed the study. M.G. established protocols for generation of translocation and insertion intermediates, processed and interpreted the cryo-electron microscopy structures, M.G and T.B built molecular models for the Sec61 complex and prepared all figures. B.B. prepared SRP and assisted in preparing wheat germ translation extract and PKRM. J.-P.A. assisted in data processing, C.B.-G. built a refined model of the wheat germ ribosome. O.B. performed cryo-electron microscopy data collection, M.G., T.B. and R.B. interpreted results and wrote the paper.

Author Information Cryo-electron microscopy maps for the idle 80S–Sec61 complex, the LepT–RNC–Sec61 complex and the LepM–RNC–Sec61 complex have been deposited in the EMDataBank with accession codes EMD-2510, EMD-2511 and EMD-2512. The respective coordinates for electron-microscopy-based models are deposited in the Protein Data Bank (4CG7, 4CG5 and 4CG6 for the Sec61 complex and 3J5Z, 3J60, 3J62, 3J61 for the updated model of the wheat germ ribosome). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.B. (beckmann@lmb.uni-muenchen.de).

LETTER

doi:10.1038/nature12824

Aprataxin resolves adenylated RNA–DNA junctions to maintain genome integrity

Percy Tumbale[1]*, Jessica S. Williams[1,2]*, Matthew J. Schellenberg[1]*, Thomas A. Kunkel[1,2] & R. Scott Williams[1]

Faithful maintenance and propagation of eukaryotic genomes is ensured by three-step DNA ligation reactions used by ATP-dependent DNA ligases[1,2]. Paradoxically, when DNA ligases encounter nicked DNA structures with abnormal DNA termini, DNA ligase catalytic activity can generate and/or exacerbate DNA damage through abortive ligation that produces chemically adducted, toxic 5′-adenylated (5′-AMP) DNA lesions[3–6]. Aprataxin (APTX) reverses DNA adenylation but the context for deadenylation repair is unclear. Here we examine the importance of APTX to RNase-H2-dependent excision repair (RER) of a lesion that is very frequently introduced into DNA, a ribonucleotide. We show that ligases generate adenylated 5′ ends containing a ribose characteristic of RNase H2 incision. APTX efficiently repairs adenylated RNA–DNA, and acting in an RNA–DNA damage response (RDDR), promotes cellular survival and prevents S-phase checkpoint activation in budding yeast undergoing RER. Structure–function studies of human APTX–RNA–DNA–AMP–Zn complexes define a mechanism for detecting and reversing adenylation at RNA–DNA junctions. This involves A-form RNA binding, proper protein folding and conformational changes, all of which are affected by heritable *APTX* mutations in ataxia with oculomotor apraxia 1. Together, these results indicate that accumulation of adenylated RNA–DNA may contribute to neurological disease.

Previous studies indicate that abortive ligation (Fig. 1a) may occur during attempts to repair DNA lesions generated by oxidation[4–7] or alkylation[7,8]. We explored a much more abundant opportunity for abortive ligation, that is, during ribonucleotide excision repair (RER)[9–11]. RER is initiated when RNase H2 cleaves on the 5′ side of a ribonucleotide found in a 5′-RNA–DNA-3′ junction (Fig. 1b, referred to hereafter as RNA–DNA junction). This event is estimated to generate more than 1,000,000 nicked RNA–DNA junctions per cell cycle in mice[10] and more than 10,000 nicked RNA–DNA junctions per cell cycle in budding yeast[11–13]. Our study was prompted by the fact that ribonucleotides are introduced into the nuclear genome at levels that are much greater than all known types of DNA damage combined, and evidence that DNA ligation *in vitro* is impaired at incised RNA–DNA junctions[1,14]. We compared the ability of human DNA ligase I to seal a nick containing canonical 3′-OH and 5′-P termini to a nick containing a 3′-OH and a 5′-P attached to a rG that mimics a nick generated when RNase H2 initiates RER (Fig. 1b). Greater than 95% of the nicked DNA substrate containing the 3′-OH and 5′-P termini was ligated within 10 min. In contrast, the presence of a single ribonucleotide (rG) on the 5′ side of the nick (5′ RNA substrate, Fig. 1b) significantly impaired generation of the 39-nucleotide ligation product (<1% ligation at 10 min, Extended Data Fig. 1a). Ligase I processing of the 5′ RNA substrate also produced an additional species migrating at a size of ~20 nucleotides, that corresponds to a bona fide 5′-adenylated product (5′-AMP$^{\mathrm{RNA-DNA}}$) (Fig. 1b and Extended Data Fig. 1b). The adenylated product comprises greater than 50% of all DNA ligase I catalytic events on the 5′ RNA substrate at all time points measured (Fig. 1c and Extended Data Fig. 1a). Also, human DNA ligase III and bacteriophage T4 DNA ligase, but not *Escherichia coli* NAD-dependent LigA, generated similar amounts of ribonucleotide-triggered abortive ligation products (Fig. 1c). Thus, incised RNA–DNA junctions

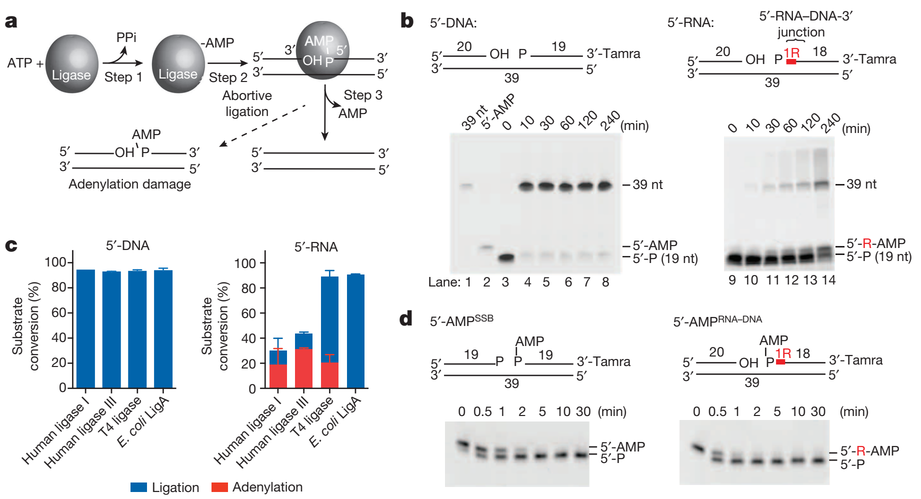


Figure 1 | Abortive ligation at RNA–DNA junctions is resolved by APTX. **a**, ATP-dependent DNA ligation: (1), ATP-dependent DNA ligase adenylation; (2), AMP is transferred to the DNA 5′ phosphate to form 5′-AMP; (3), alignment of a DNA 3′-hydroxyl with 5′-AMP within the ligase active site facilitates the nick-sealing reaction. Ligase encounter with distorting termini triggers abortive ligation. **b**, DNA ligation is aborted at RNA–DNA junctions. The red 'R' indicates the position of ribonucleotide. **c**, Quantification of total catalytic events producing sealed DNA ends (39-nucleotide product, blue bars) or abortive DNA adenylation (red bars) by DNA ligases. Mean ± s.d. ($n = 2$ replicates) is displayed for 60-min ligation reactions. **d**, Human APTX DNA-adenylate hydrolysis. Reactions contained 2 nM human APTX and 10 nM of the indicated substrate.

[1]Laboratory of Structural Biology, National Institute of Environmental Health Sciences, NIH, DHHS, Research Triangle Park, North Carolina 27709, USA. [2]Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, NIH, DHHS, Research Triangle Park, North Carolina 27709, USA.
*These authors contributed equally to this work.

are poor substrates for eukaryotic DNA ligase nick-sealing reactions, and also trigger abortive ligation at high frequency *in vitro*.

Aprataxin deadenylase (APTX in mammals and *Schizosaccharomyces pombe*, and Hnt3 in *Saccharomyces cerevisiae*) reverses DNA adenylation[3–5]. Inactivation of *APTX* in ataxia oculomotor apraxia 1 (AOA1)[15–17] suggests that persistent adenylated DNA strand breaks drive cerebellar degeneration in neurological disease[4]. However, the molecular context for APTX deadenylation remains uncertain. To examine a potential role for APTX during RER, we compared steady-state kinetic parameters for deadenylation by human APTX on gel-purified abortive ligation substrates arising from metabolism of RNA–DNA junctions (5′-AMP$^{\text{RNA–DNA}}$) to those representative of abortive ligation on DNA single-strand breaks created by reactive oxygen species[4] (5′-AMP$^{\text{SSB}}$) (Fig. 1d and Extended Data Fig. 1c). Both substrates were efficiently processed with comparable rates ($k_{cat} = 0.31$ versus $0.37\ s^{-1}$) with catalytic efficiencies that are ~30,000-fold higher than those reported on nucleotide substrates[18]. A ~6-fold higher k_{cat}/K_m for 5′-AMP$^{\text{RNA–DNA}}$ versus 5′-AMP$^{\text{SSB}}$ indicates that human APTX displays an *in vitro* preference for the RNA–DNA-derived substrates.

Both *S. pombe* Aptx and *S. cerevisiae* Hnt3$^{\text{Aptx}}$ also harbour 5′-AMP$^{\text{RNA–DNA}}$ deadenylase activity (Extended Data Fig. 1d, e). To determine whether Aptx deadenylates abortive ligation products generated at RNA–DNA junctions *in vivo*, we examined whether the phenotypes of budding yeast strains with varying capacity to incorporate and repair ribonucleotides were altered by Hnt3$^{\text{Aptx}}$ deficiency (Fig. 2). A M644G variant of the leading strand replicase, DNA polymerase ε (Pol ε, encoded by the *POL2* gene, see Extended Data Table 1), has increased capacity to incorporate ribonucleotides into DNA *in vitro* and *in vivo*[13,19]. We generated heterozygous diploids in which one copy of *HNT3* was replaced with the *NatMX4* marker. Tetrad analysis showed that although *HNT3* is dispensable for growth in a wild-type Pol ε (*POL2*) strain, growth of *pol2-M644G hnt3Δ* haploids is severely impaired (Fig. 2a), with macroscopic colonies only observed after extended incubation (Extended Data Fig. 2a).

We reasoned that the growth defect of the *pol2-M644G hnt3Δ* strain is linked to accumulation of persistent adenylated DNA strand breaks generated by DNA ligase processing of RNase H2 incised RNA–DNA junctions (that is, 5′-AMP$^{\text{RNA–DNA}}$). To test whether RNase H2 activity contributes to the impaired growth of the *pol2-M644G hnt3* mutant, we sporulated and dissected a diploid strain homozygous for deletion of the gene encoding the catalytic subunit of RNase H2 (*RNH201*). Notably, deleting *RNH201* (*rnh201Δ*) largely mitigated the growth defect of the *pol2-M644G hnt3Δ* mutant (Fig. 2b, c). This observation indicates that incision of ribonucleotides in DNA by RNase H2 generates an RER intermediate leading to production of 5′-AMP$^{\text{RNA–DNA}}$ that requires deadenylation by Hnt3$^{\text{Aptx}}$ (see model, Fig. 2e).

Increased Rnr3 protein level is a sensitive indicator of S-phase checkpoint activation[12,20]. An increased level of the Rnr3 subunit of ribonucleotide reductase was detected in *pol2-M644G hnt3Δ* cells (Fig. 2d, lane 6), but was reduced in the triple mutant *pol2-M644G hnt3Δ rnh201Δ* strain (lane 8) to a level equivalent to that of a *pol2-M644G rnh201Δ* mutant (lane 7). This suggests that failure of Hnt3$^{\text{Aptx}}$ to deadenylate 5′-AMP$^{\text{RNA–DNA}}$ lesions activates the S-phase checkpoint. We also tested *hnt3Δ* mutant strains for sensitivity to genotoxic stress caused by hydroxyurea (HU). HU treatment increases rNMP incorporation[10] and induces replication fork stalling. Growth of the *pol2-M644G hnt3Δ* mutant on rich medium was slowed, and survival in the presence of HU was reduced (Extended Data Fig. 2b, c). Notably, deleting *RNH201* reduced HU sensitivity to a level comparable to *pol2-M644G rnh201Δ* cells (Extended Data Fig. 2c).

Next we examined the consequences of loss of Hnt3 function in yeast strains containing a Pol ε variant with reduced capacity to incorporate ribonucleotides, *pol2-M644L* (ref. 13). With fewer ribonucleotides in the genome, the *pol2-M644L hnt3Δ* mutant displayed normal growth (Fig. 2c) and was unaffected by deleting *RNH201*. The stark contrast between the consequences of loss of Hnt3 function in the *pol2-M644G* variant (high genomic ribonucleotides) versus the *pol2-M644L* mutant (reduced genomic ribonucleotides) is consistent with the model wherein Hnt3$^{\text{Aptx}}$ deadenylates genotoxic abortive ligation intermediates arising during RER of ribonucleotides incorporated by Pol ε during DNA replication (Fig. 2e). A genetic interaction between *HNT3* and *RNH201* is not apparent in a *POL2* strain, possibly because adenylated RNA–DNA junctions may be removed by alternative nucleolytic processing, for example, mediated by Rad27$^{\text{Fen1}}$ and Mre11/Rad50/Xrs2$^{\text{Nbs1}}$ nucleases[8].

Having implicated aprataxins in processing 5′-AMP$^{\text{RNA–DNA}}$ *in vitro* and *in vivo*, we aimed to define the molecular basis for 5′-AMP$^{\text{RNA–DNA}}$ processing by human APTX. Structural analysis of the *S. pombe* APTX DNA complex revealed the architecture of the yeast Aptx HIT-Znf domain, and a basis for engagement of DNA ends[3]. However, the molecular basis for the APTX RNA–DNA interactions, and the mechanism of the APTX DNA damage direct reversal catalytic reaction, remain

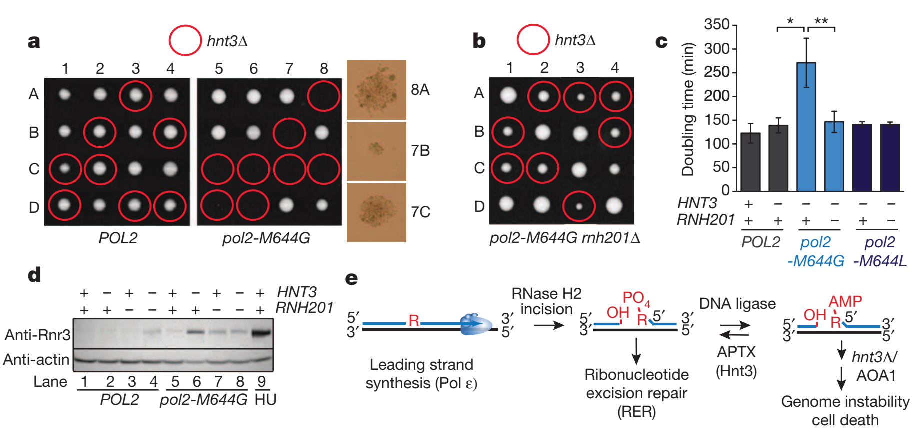


Figure 2 | Yeast Hnt3$^{\text{Aptx}}$ is critical for resolving abortive ligation intermediates that arise after incision at genomic ribonucleotides by RNase H2. **a**, Tetrad analysis of *HNT3/hnt3::natMX* diploids. 1–8 are tetrad dissections and A–D are haploid spore colonies. Right: day 3 microscopic spore colonies in the *pol2-M644G hnt3Δ* strains. **b**, Tetrad analysis of *HNT3/hnt3::natMX* diploids in the *pol2-M644G rnh201Δ* background. Plates imaged at 3 days. **c**, Deletion of *HNT3* in the *pol2-M644G* mutator confers a slow growth phenotype that is eliminated by deleting *RNH201*. Doubling times (D_t) were calculated from cultures in the logarithmic phase of growth in rich medium at 30 °C. Average doubling time ± s.d. are calculated from four biological replicates (eight for the *pol2-M644G hnt3Δ* genotype, $*P < 0.0007$; $**P < 0.0011$ (two-tailed *t*-test)). **d**, Immunoblotting of whole-cell extracts was performed using an antibody to Rnr3. **e**, RNase H2 cleavage at ribonucleotides incorporated during Pol ε leading-strand DNA synthesis leads to abortive ligation intermediates requiring APTX processing. Deletion of *HNT3* (*hnt3Δ*) or APTX deficiency in ataxia oculomotor apraxia 1 (AOA1) creates persistent adenylated strand breaks.

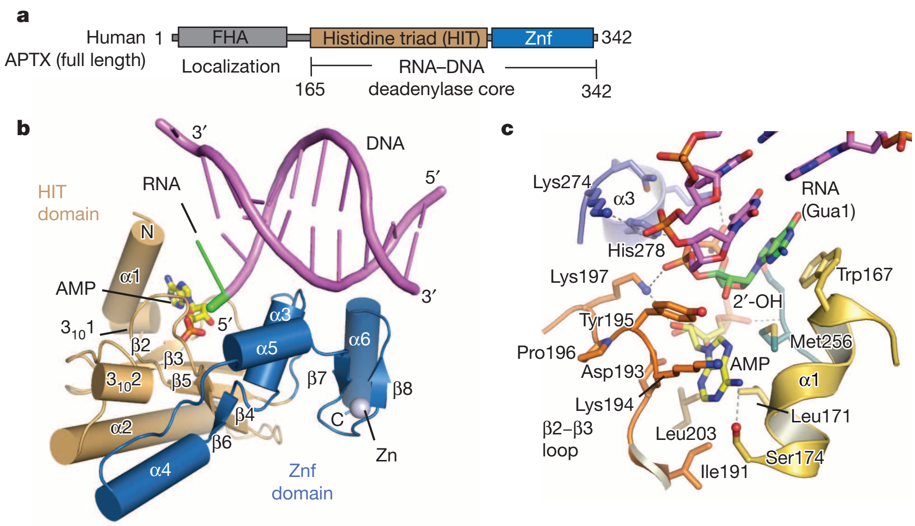


Figure 3 | Recognition of adenylated RNA–DNA junctions by human APTX. **a**, Domain architecture of human APTX. The RNA deadenylase core used for structural studies maps to residues 165–342. **b**, X-ray structure of the human APTX–RNA–DNA–AMP–Zn reaction product complex. The APTX HIT domain (tan) and Znf domain (blue) are displayed as cartoon-representation helices (cylinders) and β-strands. DNA is displayed as magenta duplex, with a green 5′-ribonucleotide and yellow AMP lesions. **c**, Four conserved elements dictate interactions with the 5′-ribonucleotide (green) and 5′-AMP (yellow with orange/red phosphate group). The β2-β3-loop (orange), HIT α1 (gold), Znf α3 (blue) and HΦHΦH loop (dark green) completely envelop and orient the 5′-adenylated ribonucleotide lesion for catalytic processing.

unclear (see Supplementary Discussion). The minimal catalytic domain of human APTX was mapped to residues 165–342 using deletion mutagenesis, limited proteolysis and deadenylation assays (Extended Data Fig. 3a–c). We then determined four X-ray crystal structures of: (1) an RNA–DNA-bound human APTX–5′-AMP–RNA–DNA–Zn quaternary product complex; (2) a mimic of an adenylated RNA–DNA processing enzymatic transition state; (3) a DNA-only bound human APTX–5′-AMP–DNA–Zn quaternary complex structure; and (4) an AOA1 mutant human APTX(K197Q) RNA–DNA bound quaternary product complex (see Supplementary Discussion, Extended Data Table 2 and Extended Data Fig. 3).

The APTX α-β histidine triad (HIT) fold domain[21] assembles with a DNA-binding Znf domain in human APTX RNA–DNA deadenylase (Fig. 3b and Extended Data Figs 4 and 5). Close interactions between the HIT and Znf subdomains mould both the active site and the extended RNA–DNA damage-binding surface (Fig. 3b and Extended Data Fig. 5). The 5′-adenylate binding pocket and 5′-ribonucleotide interaction surfaces localize to the intersection of the HIT and Znf domains (Fig. 3b, c). The APTX-bound RNA–DNA junction is significantly distorted from B-form DNA (Extended Data Fig. 5a). A two-point nucleic acid–protein interaction induces a ~15° bend in the RNA–DNA by anchoring the exposed 5′-terminal RNA base stack and the 5′-AMP lesion on one side, while engaging the opposite undamaged strand with an array of contacts from the Znf domain (Extended Data Fig. 5d–g). Biochemical studies revealed that APTX disrupts Watson–Crick base pairing of the adenylated base pair[22]. In our structures, DNA distortions and capping of the RNA–DNA base-stack by the HIT domain amino-terminal helix (α1) provide a possible mechanism for un-pairing of the terminal rG•C base pair to gain access to the lesion (Extended Data Fig. 5b–g). In the DNA-only bound human APTX structure, similar DNA distortions are observed, revealing that APTX processes adenylated RNA–DNA and DNA with an analogous mode of substrate engagement (Extended Data Fig. 6 and Supplementary Discussion). APTX sequesters the 5′-AMP lesion into a hydrophobic active site recess in an extra-helical conformation that is rotated ~180° relative to the RNA–DNA helical axis (Fig. 3c and Extended Data Fig. 5c).

RNA–DNA damage detection and reaction chemistry are mediated by four stringently conserved APTX elements that converge on the 5′-ribonucleotide and 5′-AMP lesion: HIT helix α1, the 'histidine triad' HΦHΦH loop (where H is histidine and Φ denotes a hydrophobic amino acid), the 'β2-β3' loop, and Znf helix α3 (Fig. 3c and Extended Data Fig. 4). The two 5′-terminal nucleotides of the damaged strand are bound in an A-form conformation, consistent with an RNA–DNA processing role for the aprataxins. Multiple contacts bind a C3′-endo sugar-puckered 5′-rG, including ribose sugar-phosphate interactions from Tyr 195 and Lys 197 of the β2-β3 loop, and aromatic base stacking from Trp 167 of HIT α1 (Fig. 3c and Extended Data Figs 3f and 5c). Cradling of the 2′-hydroxyl of the ribonucleotide with van der Waals interactions from Tyr 195 (β2-β3 loop) and Met 256 of the HΦHΦH loop further anchors the 5′-rG and aids in aligning the 5′-adenylated RNA terminus for catalysis (Fig. 3c and Extended Data Fig. 5c). Mutational studies underscore the importance of the β2-β3 loop in substrate binding and catalytic activity (Supplementary Discussion and Extended Data Fig. 6).

The first step of the APTX reaction is proposed to generate a covalent enzyme–AMP intermediate[3,22], via an enzyme–nucleic acid transition state that poses a significant challenge to protein structural interrogation. To trap this transition state, we developed reaction conditions under which APTX activity is inhibited when co-incubated with adenosine, orthovanadate and a 5′-phosphorylated RNA–DNA junction duplex (Extended Data Fig. 3g, h). Reaction of human APTX with these reagents *in crystallo* produced a mimic of the enzyme–RNA–DNA–AMP transition state intermediate for step 1 of a two-step deadenylation reaction (Fig. 4a and Extended Data Fig. 3i, j).

The HΦHΦH loop completely encircles the adenylated 5′-ribonucleotide lesion (Fig. 4), with His 260 covalently bonded to a pentavalent coordinate vanadium atom in the transition state mimic complex (Fig. 4a, d and Extended Data Fig. 3i). The transition state and product-bound structures support a two-step deadenylation reaction initiated by nucleophilic attack of the scissile pyrophosphate by His 260 (Fig. 4d). Protein main-chain amides of Ser 255-Met 256 and salt bridging from His 201–His 262 stabilize this transition state (Fig. 4a). His 251 is ideally positioned to protonate a 5′-P leaving group, and binds 5′-P together with Ser 255 and Lys 277. In the proposed reaction scheme (Fig. 4d), step 1 generates an enzyme–AMP intermediate, which is then resolved via hydrolysis in step 2. Notably, in the product complex, a Na^+ ion (assigned by oxygen-ion bond lengths) with octahedral coordination binds between Ser 255 and 5′-P (Fig. 4b), indicating that although APTX activity is metal-independent, transient solvent cation binding stabilizes the product state, similar to third metal binding in DNA polymerase η[23].

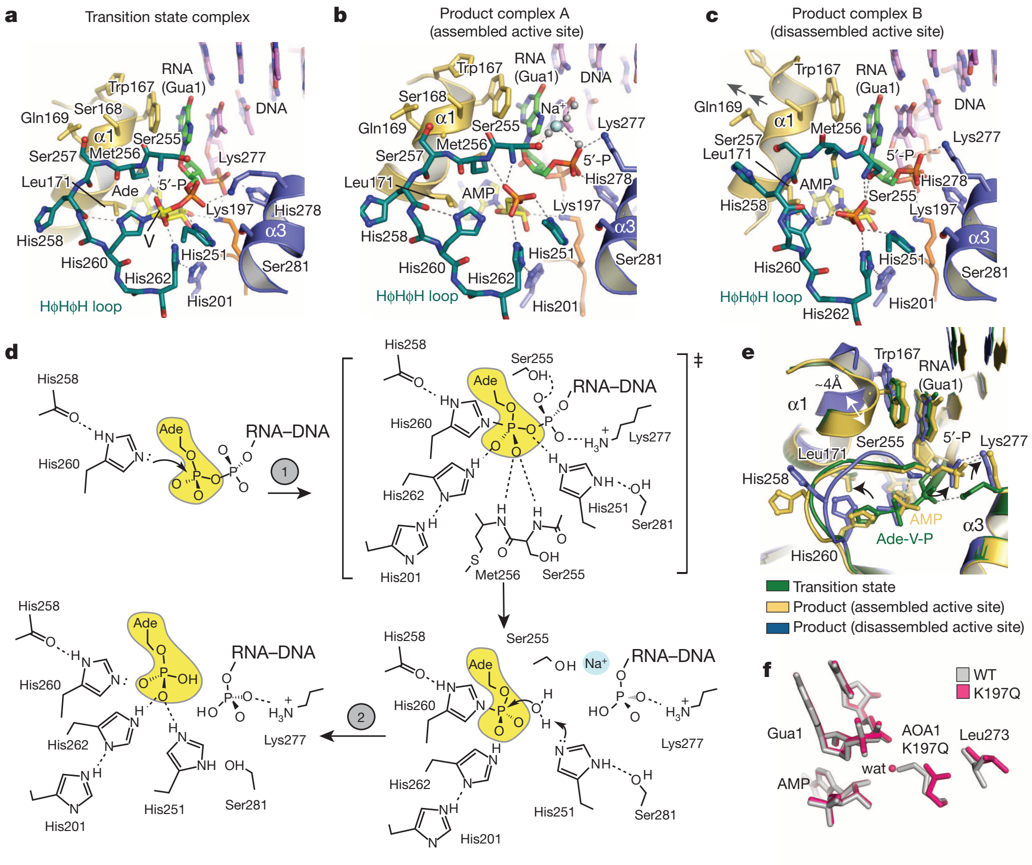


Figure 4 | RNA–DNA deadenylation reaction mechanism and APTX inactivation in AOA1. **a**, Human APTX–RNA–DNA–adenosine–vanadate transition state mimic complex active site. **b**, Product complex (assembled active site) with 5′-ribonucleotide (green) and 5′-AMP (yellow) bound in the substrate interaction cleft. **c**, Product complex (disassembled active site) substrate interaction cleft. **d**, Proposed human APTX reaction mechanism. **e**, Structural overlays of human APTX states illustrate the coupled movement of the N-terminal α1 helix and the HΦHΦH active-site loop. **f**, Structural repercussions of the AOA1 APTX(K197Q) variant. A structural overlay of wild type (grey) and mutant K197Q (pink).

Human APTX is found in two markedly different conformations in the product-bound structure. The first conformation (the ‘assembled active site’ Fig. 4b and Extended Data Fig. 7a, b) has an intact active site characterized by close interactions between HIT α1 (Leu 171 and Trp 167) and the HΦHΦH loop, and correct positioning of His 260 for catalysis. This state has the His 260 imidazole ring hydrogen bonded to the His 268 main-chain carbonyl oxygen. In the second state (the ‘disassembled active site’), α1 is displaced by ~4 Å relative to a rearranged HΦHΦH loop, and His 260 is flipped out of alignment for nucleophilic attack (compare Fig. 4b and c). Structural overlays (Fig. 4e) and interpolations between these two states (Supplementary Videos 1 and 2) indicate that concerted conformational rearrangements sculpt the HΦHΦH loop, and may be linked to RNA–DNA substrate binding by α1 and HΦHΦH (Extended data Fig. 7b–e). We propose that interactions between RNA–DNA and protein proximal to the active site regulate active-site conformations involving HIT α1. RNA/DNA-regulated assembly of the APTX active site may ‘license’ catalytic activity and also prevent inappropriate, nonspecific hydrolysis of nucleotides (for example, ATP or ADP hydrolysis). Discrimination against ATP cleavage may be critical for mitochondrial APTX isoforms that have previously been implicated in DNA damage repair in mitochondria[24], because off-target catalysis could imbalance nucleotide pool levels.

Both missense and truncating APTX substitutions are linked to neurodegenerative disease[15–17]. On the basis of the human APTX structures determined here, we predict that most AOA1 mutations (D185E, A198V, P206L, G231E, R247X, V263G, D267G, W279X, W279R and R306X) will decrease protein stability by truncating the polypeptide or by altering the protein-folding core (Extended Data Fig. 8a). Conformational differences between our RNA–DNA bound structures extend into the protein core (Extended Data Fig. 7a). APTX conformational changes may thus be subject to mutagenic modulation in disease. We posit that differential impacts on protein folding, active-site chemistry and substrate induced-fit active site assembly may all contribute to the variable clinical outcomes observed in patients with APTX defects[17].

One AOA1 mutation is found in the RNA–DNA substrate interaction cleft (K197Q) and two participate directly in active-site chemistry (H201R and H201Q) (Fig. 4a–c and Extended Data Fig. 8a). The late-onset AOA1 variant APTX(K197Q)[17] displays significantly impaired deadenylation activity on both the 5′-AMP^{SSB} and 5′-AMP$^{RNA-DNA}$ substrates (Extended Data Fig. 6b). To understand the molecular basis for the K197Q defect, we determined a 1.90 Å X-ray structure of APTX(K197Q) bound to RNA–DNA and AMP that reveals the mutant protein harbours a distorted active-site pocket (Fig. 4f and Extended Data Fig. 8b, c). In the wild-type protein, Lys 197 participates in salt-bridging interactions with the 5′-terminal sugar-phosphate backbone and the AMP lesion 2′-hydroxyl. In the mutant, Gln 197 is rotated away from the substrate-binding pocket and substitutes direct protein–substrate interaction with a protein–water–substrate nucleic acid binding interaction, thus revealing that distortions in the APTX(K197Q) substrate-binding pocket underlie AOA1.

Our data indicate that during repair of non-canonical ribonucleotides introduced into DNA during replication of the nuclear genome, DNA ligases generate 5′-adenylated RNA–DNA junctions that can elicit a DNA damage checkpoint response unless this is prevented by APTX deadenylase. In addition to frequent ribonucleotide incorporation by DNA replicases, rNTPs are used by RNA primase to initially

synthesize ~5% of the nascent lagging strand, and rNTPs are also incorporated during mitochondrial DNA replication[25,26], during translesion synthesis[27], and during DNA repair[28]. Ribonucleotide incorporation during DNA repair may be more prevalent in non-proliferating cells because dNTP concentrations are lower[29,30], thereby increasing rNTP:dNTP ratios[30]. Thus, the late onset of AOA1 might partly reflect failure to deadenylate RNA–DNA junctions resulting from ribonucleotides incorporated in DNA transactions occurring over many years in quiescent neurons. It will be important in future work to establish quantitative measures of RNA–DNA adenylation to explore this hypothesis.

In this context, APTX acts in a nucleic acid transaction that is not exclusively DNA or RNA. Instead, using a reaction mechanism that is finely tuned to operate on RNA–DNA junctions, APTX acts in an RNA–DNA damage response (RDDR) to protect the genome from a compound insult, a ribosylated, adenylated 5′ terminus. In a broader sense, it seems probable that other enzymes may also modulate the RDDR via the detection, processing and signalling of RNA–DNA-derived structures posing threats to genomic integrity.

METHODS SUMMARY

Proteins were expressed in *Escherichia coli* and purified with standard procedures. All crystals were grown using sitting-drop vapour diffusion. X-ray diffraction data were all collected at 100 K at the Advanced Photon Source, beamlines 22-ID and 22-BM. Initial DNA-bound human APTX structures were solved by molecular replacement with the *S. pombe* Aptx–DNA complex (RCSB code 3SZQ). RNA–DNA-bound wild-type and mutant human APTX structures were solved by molecular replacement using the refined human APTX–DNA-bound model. *S. cerevisiae* strain construction and growth assays were performed as described[12].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 April; accepted 31 October 2013.

Published online 22 December 2013.

1. Pascal, J. M., O'Brien, P. J., Tomkinson, A. E. & Ellenberger, T. Human DNA ligase I completely encircles and partially unwinds nicked DNA. *Nature* **432,** 473–478 (2004).
2. Ellenberger, T. & Tomkinson, A. E. Eukaryotic DNA ligases: structural and functional insights. *Annu. Rev. Biochem.* **77,** 313–338 (2008).
3. Tumbale, P. *et al.* Structure of an aprataxin-DNA complex with insights into AOA1 neurodegenerative disease. *Nature Struct. Mol. Biol.* **18,** 1189–1195 (2011).
4. Ahel, I. *et al.* The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates. *Nature* **443,** 713–716 (2006).
5. Rass, U., Ahel, I. & West, S. C. Actions of aprataxin in multiple DNA repair pathways. *J. Biol. Chem.* **282,** 9469–9474 (2007).
6. Harris, J. L. *et al.* Aprataxin, poly-ADP ribose polymerase 1 (PARP-1) and apurinic endonuclease 1 (APE1) function together to protect the genome against oxidative damage. *Hum. Mol. Genet.* **18,** 4102–4117 (2009).
7. El-Khamisy, S. F. *et al.* Synergistic decrease of DNA single-strand break repair rates in mouse neural cells lacking both Tdp1 and aprataxin. *DNA Repair* **8,** 760–766 (2009).
8. Daley, J. M., Wilson, T. E. & Ramotar, D. Genetic interactions between HNT3/Aprataxin and RAD27/FEN1 suggest parallel pathways for 5′ end processing during base excision repair. *DNA Repair* **9,** 690–699 (2010).
9. Sparks, J. L. *et al.* RNase H2-initiated ribonucleotide excision repair. *Mol. Cell* **47,** 980–986 (2012).
10. Reijns, M. A. *et al.* Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. *Cell* **149,** 1008–1022 (2012).
11. Nick McElhinny, S. A. *et al.* Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases. *Proc. Natl Acad. Sci. USA* **107,** 4949–4954 (2010).
12. Williams, J. S. *et al.* Topoisomerase 1-mediated removal of ribonucleotides from nascent leading-strand DNA. *Mol. Cell* **49,** 1010–1015 (2013).
13. Nick McElhinny, S. A. *et al.* Genome instability due to ribonucleotide incorporation into DNA. *Nature Chem. Biol.* **6,** 774–781 (2010).
14. Rumbaugh, J. A., Murante, R. S., Shi, S. & Bambara, R. A. Creation and removal of embedded ribonucleotides in chromosomal DNA during mammalian Okazaki fragment processing. *J. Biol. Chem.* **272,** 22591–22599 (1997).
15. Date, H. *et al.* Early-onset ataxia with ocular motor apraxia and hypoalbuminemia is caused by mutations in a new HIT superfamily gene. *Nature Genet.* **29,** 184–188 (2001).
16. Moreira, M. C. *et al.* The gene mutated in ataxia-ocular apraxia 1 encodes the new HIT/Zn-finger protein aprataxin. *Nature Genet.* **29,** 189–193 (2001).
17. Tranchant, C., Fleury, M., Moreira, M. C., Koenig, M. & Warter, J. M. Phenotypic variability of aprataxin gene mutations. *Neurology* **60,** 868–870 (2003).
18. Kijas, A. W., Harris, J. L., Harris, J. M. & Lavin, M. F. Aprataxin forms a discrete branch in the HIT (histidine triad) superfamily of proteins with both DNA/RNA binding and nucleotide hydrolase activities. *J. Biol. Chem.* **281,** 13939–13948 (2006).
19. Lujan, S. A. *et al.* Mismatch repair balances leading and lagging strand DNA replication fidelity. *PLoS Genet.* **8,** e1003016 (2012).
20. Davidson, M. B. *et al.* Endogenous DNA replication stress results in expansion of dNTP pools and a mutator phenotype. *EMBO J.* **31,** 895–907 (2012).
21. Lima, C. D., Klein, M. G. & Hendrickson, W. A. Structure-based analysis of catalysis and substrate definition in the HIT protein family. *Science* **278,** 286–290 (1997).
22. Rass, U., Ahel, I. & West, S. C. Molecular mechanism of DNA deadenylation by the neurological disease protein aprataxin. *J. Biol. Chem.* **283,** 33994–34001 (2008).
23. Nakamura, T., Zhao, Y., Yamagata, Y., Hua, Y. J. & Yang, W. Watching DNA polymerase eta make a phosphodiester bond. *Nature* **487,** 196–201 (2012).
24. Sykora, P., Croteau, D. L., Bohr, V. A. & Wilson, D. M. III. Aprataxin localizes to mitochondria and preserves mitochondrial function. *Proc. Natl Acad. Sci. USA* **108,** 7437–7442 (2011).
25. Kasiviswanathan, R. & Copeland, W. C. Ribonucleotide discrimination and reverse transcription by the human mitochondrial DNA polymerase. *J. Biol. Chem.* **286,** 31490–31500 (2011).
26. Yang, M. Y. *et al.* Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication. *Cell* **111,** 495–505 (2002).
27. McDonald, J. P., Vaisman, A., Kuban, W., Goodman, M. F. & Woodgate, R. Mechanisms employed by *Escherichia coli* to prevent ribonucleotide incorporation into genomic DNA by Pol V. *PLoS Genet.* **8,** e1003030 (2012).
28. Nick McElhinny, S. A. & Ramsden, D. A. Polymerase mu is a DNA-directed DNA/RNA polymerase. *Mol. Cell. Biol.* **23,** 2309–2315 (2003).
29. Chabes, A. *et al.* Survival of DNA damage in yeast directly depends on increased dNTP levels allowed by relaxed feedback inhibition of ribonucleotide reductase. *Cell* **112,** 391–401 (2003).
30. Ferraro, P., Franzolin, E., Pontarin, G., Reichard, P. & Bianchi, V. Quantitation of cellular deoxynucleoside triphosphates. *Nucleic Acids Res.* **38,** e85 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the intramural research program of the US National Institutes of Health (NIH), National Institute of Environmental Health Sciences (NIEHS) projects 1Z01ES102765 to R.S.W. and Z01ES065070 to T.A.K. X-ray diffraction data were collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-ID (or 22-BM) beamline at the Advanced Photon Source, Argonne National Laboratory. Use of the Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract number W-31-109-Eng-38. We thank L. Pedersen of the NIEHS collaborative crystallography group, and the Advanced Photon Source (APS) Southeast Regional Collaborative Access Team (SER-CAT) staff for assistance with crystallographic data collection. We thank L. Pedersen and B. Wallace for critical reading of the manuscript, J. Krahn for assistance with movies, and T. Ellenberger and P. O'Brien for DNA ligase expression vectors.

Author Contributions P.T. performed biochemical studies and crystallization. M.J.S. and R.S.W. solved and refined X-ray structures. J.S.W. performed *S. cerevisiae* experiments. All authors contributed to experimental design, data analysis and preparation of the manuscript.

Author Information Molecular coordinates and structure factors for X-ray structures reported here have been deposited in the RCSB Protein Data Bank under accession codes 4NDF (human APTX–RNA–DNA–AMP–Zn complex), 4NDG (human APTX–RNA–DNA–adenosine–vanadate–Zn complex), 4NDH (human APTX–DNA–AMP–Zn complex) and 4NDI (human APTX(K197Q)–RNA–DNA–AMP–Zn complex). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.S.W. (williamsrs@niehs.nih.gov).

doi:10.1038/nature12803

Crystal structures of the Lsm complex bound to the 3′ end sequence of U6 small nuclear RNA

Lijun Zhou[1,2]*, Jing Hang[2,3]*, Yulin Zhou[1], Ruixue Wan[3], Guifeng Lu[3], Ping Yin[2], Chuangye Yan[2,3] & Yigong Shi[1,2]

Splicing of precursor messenger RNA (pre-mRNA) in eukaryotic cells is carried out by the spliceosome[1], which consists of five small nuclear ribonucleoproteins (snRNPs) and a number of accessory factors and enzymes[2]. Each snRNP contains a ring-shaped subcomplex of seven proteins and a specific RNA molecule[2–4]. The U6 snRNP contains a unique heptameric Lsm protein complex, which specifically recognizes the U6 small nuclear RNA at its 3′ end. Here we report the crystal structures of the heptameric Lsm complex, both by itself and in complex with a 3′ fragment of U6 snRNA, at 2.8 Å resolution. Each of the seven Lsm proteins interacts with two neighbouring Lsm components to form a doughnut-shaped assembly, with the order Lsm3–2–8–4–7–5–6. The four uridine nucleotides at the 3′ end of U6 snRNA are modularly recognized by Lsm3, Lsm2, Lsm8 and Lsm4, with the uracil base specificity conferred by a highly conserved asparagine residue. The uracil base at the extreme 3′ end is sandwiched by His 36 and Arg 69 from Lsm3, through π–π and cation–π interactions, respectively. The distinctive end-recognition of U6 snRNA by the Lsm complex contrasts with RNA binding by the Sm complex in the other snRNPs. The structural features and associated biochemical analyses deepen mechanistic understanding of the U6 snRNP function in pre-mRNA splicing.

Four of the five snRNPs (U1, U2, U4 and U5) share the Sm heptamer ring[4]. By contrast, the heptamer ring in the U6 snRNP contains seven Sm-like (Lsm) proteins: Lsm2, Lsm3, Lsm4, Lsm5, Lsm6, Lsm7 and Lsm8 (refs 5–7). The U6 snRNP participates in formation of the precatalytic spliceosome and the two cleavages of pre-mRNA splicing[8–12]. U6 snRNP in yeast contains the Lsm2–8 heptamer, a 112-nucleotide RNA[13], and Prp24 (refs 14–16). The 3.6 Å resolution crystal structure of the Sm heptamer ring bound to U4 snRNA uncovered a conserved pattern of RNA recognition[17]. Another Lsm heptameric complex, Lsm1–7, which shares six components with the Lsm2–8 complex, functions in mRNA degradation pathway in the cytoplasm[18]. The recombinant Lsm heptameric complexes have been successfully reconstituted *in vitro* through denaturation and refolding[19].

We co-expressed all seven Lsm proteins from *Saccharomyces cerevisiae* and purified the heptameric Lsm2–8 complex to homogeneity (Extended Data Fig. 1a). Consistent with the finding that the Lsm2–8 complex specifically recognized a uridine-rich sequence derived from the 3′ end of U6 snRNA[5], the U6 snRNA sequences from multiple species share four consecutive uridine nucleotides at their 3′ ends (Extended Data Fig. 1b). We synthesized seven RNA oligonucleotides, each derived from the 3′ end of *S. cerevisiae* U6 snRNA, and examined their binding to the Lsm2–8 complex using isothermal titration calorimetry (ITC) (Fig. 1a). The tetra-uridine oligonucleotide 5′-$UUUU_{112}$-3′ bound to the Lsm2–8 complex with a dissociation constant of 694 ± 106 nM (Fig. 1a and Extended Data Fig. 2a). The pentanucleotide 5′-$GUUUU_{112}$-3′ had a dissociation constant of 109 ± 6 nM, a 6.4-fold increase over 5′-$UUUU_{112}$-3′. Further extension of the 5′-sequence only led to slightly enhanced binding (Fig. 1a and Extended Data Fig. 2b–g). This result demonstrates that a short RNA oligonucleotide derived from the 3′ end of U6 snRNA constitutes a minimal RNA element for specific recognition by the Lsm2–8 heptameric complex.

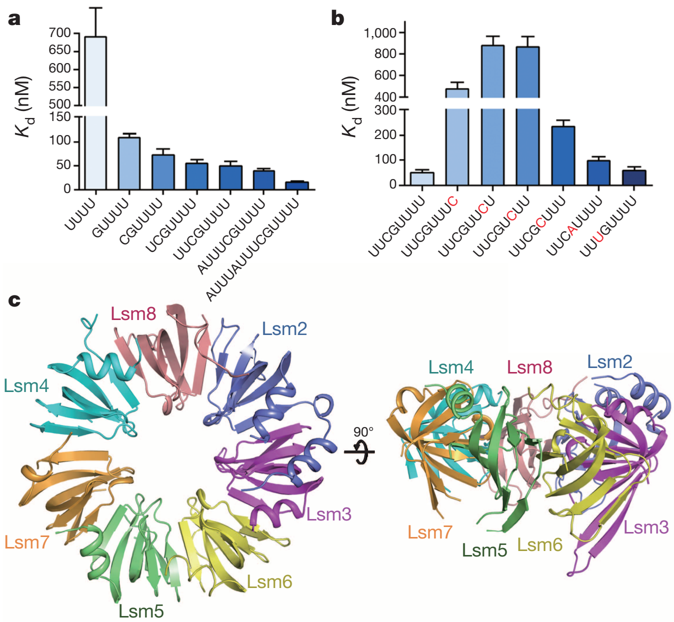


Figure 1 | Structure of the Lsm2–8 heptameric complex. **a**, Five nucleotides at the 3′ end of U6 snRNA is responsible for the bulk of binding energy between U6 snRNA and the Lsm2–8 complex. The data summarized here represent the median of three independently performed isothermal titration calorimetry (ITC) experiments. Error bars represent s.d. The same applies to Figs 1b and 4a. **b**, Mutation of any of the five nucleotides at the 3′ end of U6 snRNA results in decreased binding affinity for the Lsm2–8 complex. **c**, Overall structure of the Lsm2–8 complex in two perpendicular views.

We investigated the sequence requirement. For the octanucleotide 5′-$UUCGUUUU_{112}$-3′, each of the six bases starting from the 3′ end was replaced by a similar base — uracil by cytosine and guanine by adenine. The mutated oligonucleotides were examined for binding to the Lsm2–8 complex. Whereas the Lsm2–8 complex has a binding affinity of 52 ± 7 nM towards the wild-type (WT) octanucleotide, individual replacement of the four uracil bases at the 3′ end results in reduction of binding affinity by at least 4.5-fold (Fig. 1b and Extended Data Fig. 3a–d). Replacement of the fifth nucleotide from the 3′ end only reduced binding affinity by twofold (Fig. 1b and Extended Data Fig. 3e). Replacement of the sixth nucleotide had little effect on binding (Fig. 1b and Extended Data Fig. 3f). These results indicate that the five nucleotides from the 3′ end of U6 snRNA may represent the optimal sequence for binding by the Lsm2–8 complex.

Reasoning that extended hydrophilic sequences at the carboxy termini of Lsm4/Lsm8 (Extended Data Fig. 4a) may hinder crystal packing, we generated three Lsm complexes in which these sequences were

[1]Ministry of Education Key Laboratory of Protein Science, Tsinghua University, Beijing 100084, China. [2]Tsinghua-Peking Joint Center for Life Sciences, Center for Structural Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China. [3]State Key Laboratory of Bio-membrane and Membrane Biotechnology, Tsinghua University, Beijing 100084, China.
*These authors contributed equally to this work.

deleted. Truncation of both C termini had no significant effect on RNA binding (Extended Data Fig. 4b). We crystallized the truncated Lsm2–8 complex in two different space groups, among which $P2_1$ yielded better crystals. The structure was determined at 2.8 Å resolution by molecular replacement using the atomic coordinates of the Lsm1–7 complex (PDB code 4M75) (Extended Data Table 1 and Extended Data Fig. 5a–c).

The Lsm2–8 complex has a doughnut-shaped structure, with approximately 70 Å in outer diameter, 15 Å in inner diameter, and 45 Å in thickness (Fig. 1c). Lsm3, Lsm2, Lsm8, Lsm4, Lsm7, Lsm5 and Lsm6 interact with each other to form a closed ring, with each component only contacting two neighbouring proteins. Each Lsm protein adopts a highly conserved Sm fold[20], with a tilted β-sandwich stabilized by an amino-terminal α-helix. The β-sandwich contains five anti-parallel β-strands in one sheet and three β-strands in the other; the three-stranded β-sheet is capped on one side by the N-terminal α-helix. The other side of the three-stranded β-sheet interacts with the five-stranded β-sheet of a neighbouring Lsm protein to form a contiguous eight-stranded, anti-parallel β-sheet (Fig. 1c).

This structural organization allows the main chain carbonyl and amide groups of neighbouring Lsm proteins to form multiple intermolecular hydrogen bonds (Extended Data Fig. 5d). Despite the preponderance of these main chain hydrogen bonds, the side chains from the seven Lsm proteins contribute heavily to the specific formation of the Lsm2–8 heptameric complex. For example, Tyr 8 from Lsm8 donates a hydrogen bond to Asn 59 from Lsm2, whereas Ile 44 from Lsm5 and Val 11 from Lsm6 interact with each other through van der Waals contact (Extended Data Fig. 5e).

Next, we crystallized the Lsm2–8 complex in the presence of an octanucleotide derived from the 3′ end of U6 snRNA in the space group C2. The structure was determined by molecular replacement using atomic coordinates of the free Lsm2–8 complex (Extended Data Table 1). The RNA bases show excellent electron density (Extended Data Fig. 6a); correct assignment of the RNA sequence is confirmed by anomalous bromine (Br) signal from 5-Br-uracil of the nucleotide U_{110} (Extended Data Fig. 6b).

The RNA oligonucleotide is bound within the central hole of the Lsm2–8 ring (Fig. 2a). The four consecutive nucleotides 5′-$U_{109}U_{110}U_{111}U_{112}$-3′ are recognized by Lsm4, Lsm8, Lsm2 and Lsm3, respectively. These nucleotides follow a positively charged surface groove (Fig. 2b) and gradually veer towards one side of the Lsm ring, with the fifth nucleotide G_{108} away from the central hole and bound by Lsm7. Despite the presence of octanucleotide in the crystals (Extended Data Fig. 6c), unambiguous electron density was observed only for the five consecutive nucleotides at the 3′ end (Extended Data Fig. 6a). RNA binding only induces local conformational changes in the Lsm2–8 complex (Extended Data Fig. 7). The side chains of Phe 35 and Arg 63 in Lsm2 re-orient to sandwich uracil from U_{111}, whereas Arg 72 in Lsm4 relocates to form cation–π interactions with uracil from U_{109}. The side chain of Gln57 in Lsm7 undergoes a rotation to hydrogen bond with guanine from G_{108}.

Each of the four uracil bases at the 3′ end of the U6 snRNA is specifically recognized by a conserved pattern of interactions, involving two sequence motifs DxXxN and IRGX (Fig. 2c). The Arg residue of the IRGX motif and the middle amino acid in DxXxN sandwich an RNA base through cation–π and π–π interactions, respectively; the side chain of Asn in DxXxN and the amide nitrogen atoms of Gly-Xaa in the IRGX motif make direct hydrogen bonds to the base. Accordingly, the 3′ end uracil from U_{112} is sandwiched by the side chains of His 36 and Arg 69 from Lsm3, through cation–π and π–π interactions, respectively

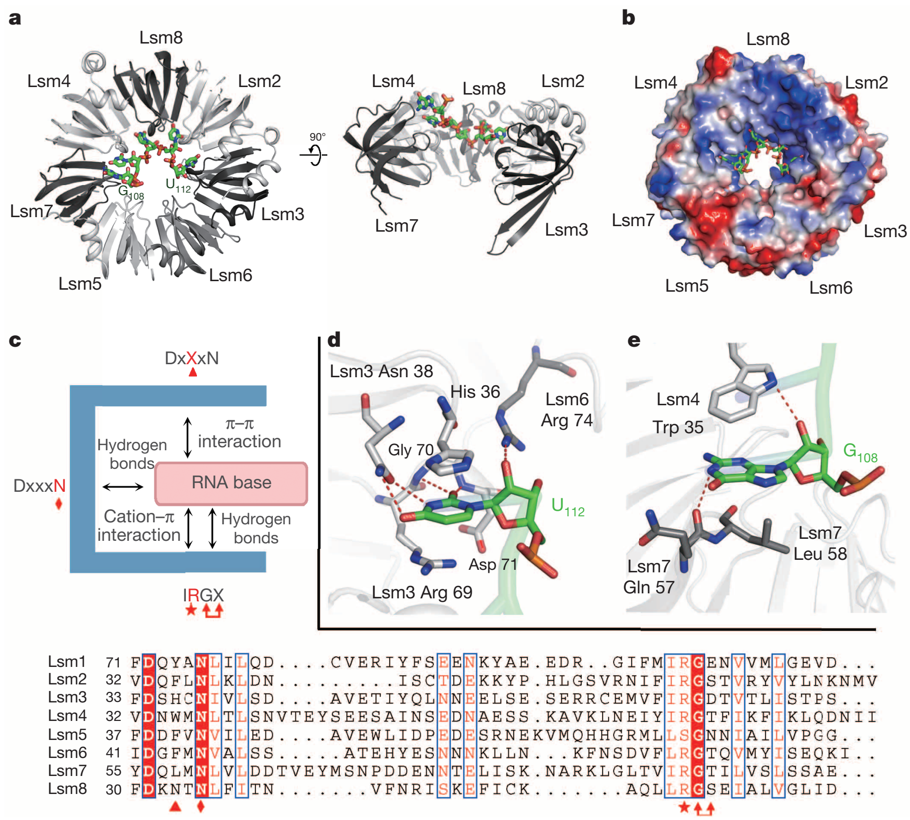


Figure 2 | Recognition of U6 snRNA by the Lsm2–8 heptameric complex. **a**, Overall structure of the Lsm2–8 complex bound to the 3′ end sequence of U6 snRNA. Five nucleotides at the 3′ end are recognized in the central hole of the Lsm ring. Two perpendicular views are shown. For clarity, Lsm5 and Lsm6 are removed in the right panel. **b**, The five nucleotides 5′-G_{108}UUUU_{112}-3′ bind to a positively charged surface region in the Lsm2–8 ring. The Lsm complex is represented by electrostatic surface potential. **c**, Schematic representation of the modular recognition of RNA bases by the Lsm proteins. The amino acids that sandwich the uracil base through π–π and cation–π interactions are indicated by red triangles and red asterisks, respectively. Uracil specificity is conferred by an invariant Asn (red diamond) and the di-residue GX of the IRGX motif (red connected arrows). **d**, Specific recognition of the nucleotide U_{112}. **e**, Coordination of the nucleotide G_{108}.

(Fig. 2d). Specific recognition of uracil is conferred by two pairs of hydrogen bonds, one pair between the 3-NH/4-O groups of uracil and the side chain of Asn 38 and the other between the 2-O atom of uracil and the main chain amide groups of Gly70 and Asp71 (Fig. 2d).

The other three uracil bases follow the same pattern of interactions (Extended Data Fig. 8). U_{111} is sandwiched between Phe 35 and Arg 63 from Lsm2, through π–π and cation–π interactions, respectively (Extended Data Fig. 8a). Notably, U_{110} no longer has the π–π interactions but maintains the conserved cation–π interactions involving Arg 57 from Lsm8 (Extended Data Fig. 8b). In addition, there are three base-specific hydrogen bonds for U_{109} or U_{110}, compared to four such hydrogen bonds for U_{111} or U_{112} (Fig. 2d and Extended Data Fig. 8). Thus, the two uracils at the 3′ end are coordinated by more interactions than the other two uracils away from the 3′ end. The gradual loosening of the interactions with U_{109} and U_{110} culminates in the departure of the fifth nucleotide G_{108} from the central hole of the Lsm ring (Fig. 2a). G_{108} is no longer coordinated by the DxXxN or the IRGX motifs, with its guanine base stacked by Trp 35 from Lsm4 and Leu 58 from Lsm7 (Fig. 2e).

The Lsm proteins share 26–40% sequence identity with the corresponding Sm proteins. Both heptameric complexes have a similar overall structure (Fig. 3a). The Lsm2–8 complex and the Sm complex recognize specific, but different, RNA elements (Fig. 3b). Recognition of individual RNA bases is quite conserved (Fig. 3c, d). The Sm proteins also contain the DxXxN and IRGX motifs and predominantly recognize uridine nucleotides. Each uracil base is sandwiched mainly by Arg and His/Phe through cation–π and π–π interactions, respectively; the base specificity is conferred through hydrogen bonds by an invariant Asn residue (Fig. 3c, d).

Despite similarity in recognition of individual bases, the overall mode of RNA recognition is quite different for the Lsm and Sm complexes. Recognition of U6 snRNA by the Lsm complex primarily involves 'end recognition', where the uridine nucleotide at the 3′ end of the RNA element is anchored by Lsm3 and the preceding three nucleotides are recognized by Lsm2/8/4 (Fig. 3b–d). By contrast, both U1 and U4 snRNAs are bound by the Sm complex through 'internal RNA recognition', where seven consecutive nucleotides are bound within the central hole of the Sm complex, each recognized by a distinct Sm protein, and the preceding and ensuing nucleotides are on two sides of the Sm ring[17,21,22] (Fig. 3b–d). Only four out of seven Lsm proteins in the Lsm2–8 complex specifically recognize the uracil bases, explaining why four consecutive uridines contribute the bulk of binding energy (Fig. 1a).

To corroborate the structural findings, we generated 21 Lsm heptameric complexes, each containing a missense mutation targeting a key residue for RNA recognition, and examined their binding to the octanucleotide 5′-UUCGUUUU-3′. Mutations in Lsm3 and Lsm2 are most deleterious, followed by mutations in Lsm4 (Fig. 4a). For these three Lsm proteins, any mutation of the conserved Asn or the two residues that sandwich the RNA base led to drastic reduction of binding affinity. By contrast, only one mutation Arg75Ala in Lsm8 resulted in similarly drastic reduction of binding affinity. Eight out of nine mutations in Lsm6/5/7 had relatively minor effect on RNA binding (Fig. 4a). The only mutation that had a pronounced effect, Arg74Ala in Lsm6, is explained by the structural observation that Arg 74 makes a hydrogen bond to the ribose of U_{112} (Fig. 2d). These biochemical observations are in excellent agreement with our structural analysis, which reveals stronger interactions for U_{112} and U_{111} than for U_{110} and U_{109} (Fig. 2d and Extended Data Fig. 8). The weakened interactions for U_{110}/U_{109} may be caused by a small degree of off-registry with respect to the 3′ end nucleotide, which accumulates over the second, third and fourth bases to disallow the fifth nucleotide G_{108} to be accommodated within the Lsm ring (Fig. 4b).

The synthetic RNA octanucleotide in our structural studies contains a 3′-OH group on the ribose of U_{112}. In cells, however, U_{112} of mature U6 snRNA is marked by a ribose 2′,3′-cyclophosphate group[23]. Presence of the cyclophosphate group was shown to slightly enhance the binding affinity for the Lsm2–8 complex[24]; this result was confirmed by our biochemical analysis (Extended Data Fig. 9). Thus the 3′ end cyclophosphate may only marginally strengthen RNA end recognition by the Lsm2–8 complex. We speculate that the negatively charged phosphate group is probably recognized by Arg 74 of Lsm6, which mediates a direct hydrogen bond to the ribose 2′-OH group of U_{112} in the crystal structure (Fig. 2d). Notably, this Arg, invariant in six Lsm proteins, is replaced by Ser in Lsm5, which borders Lsm6. This observation may explain why the 3′ end nucleotide U_{112} is coordinated by Lsm3/Lsm6, but not Lsm6/Lsm5.

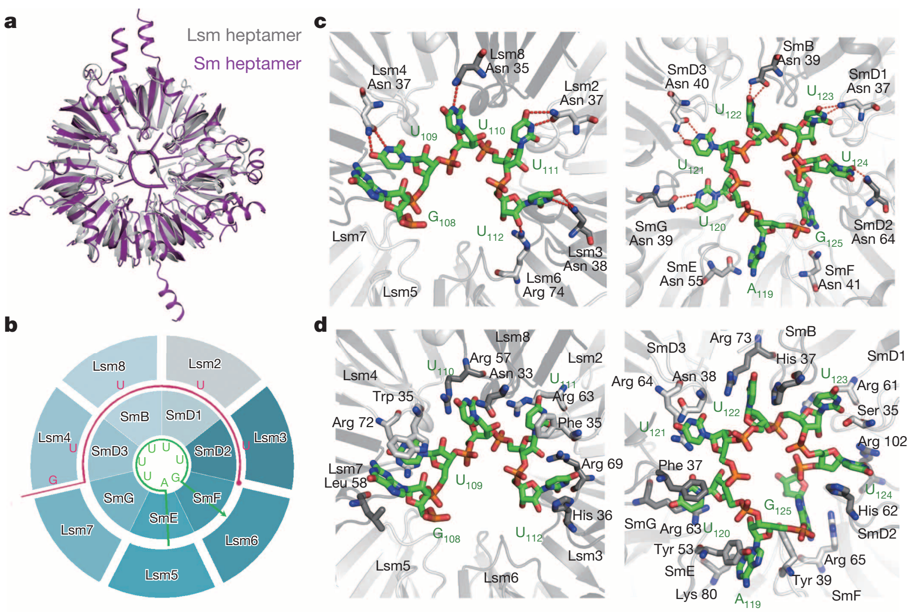


Figure 3 | Structural comparison with the Sm complex. a, The overall structure of the Lsm2–8 heptameric complex (grey) is similar to that of the Sm complex (purple). The comparison was generated by aligning Lsm8 to SmB of the Sm complex, which has a root-mean-squared deviation of 2.5 Å over 50 Cα atoms. **b**, The overall mode of RNA recognition is different between the Lsm and Sm complexes. The Lsm complex caps the 3′ end of the U6 snRNA. By contrast, the Sm complex recognizes seven consecutive nucleotides, with the preceding and ensuing nucleotides placed on two opposing sides of the Sm ring. **c**, Comparison of specific base recognition through hydrogen bonds between the Lsm complex (left panel) and the Sm complex (right panel). **d**, Comparison of base stacking interactions between the Lsm complex (left panel) and the Sm complex (right panel, PDB code 2Y9A).

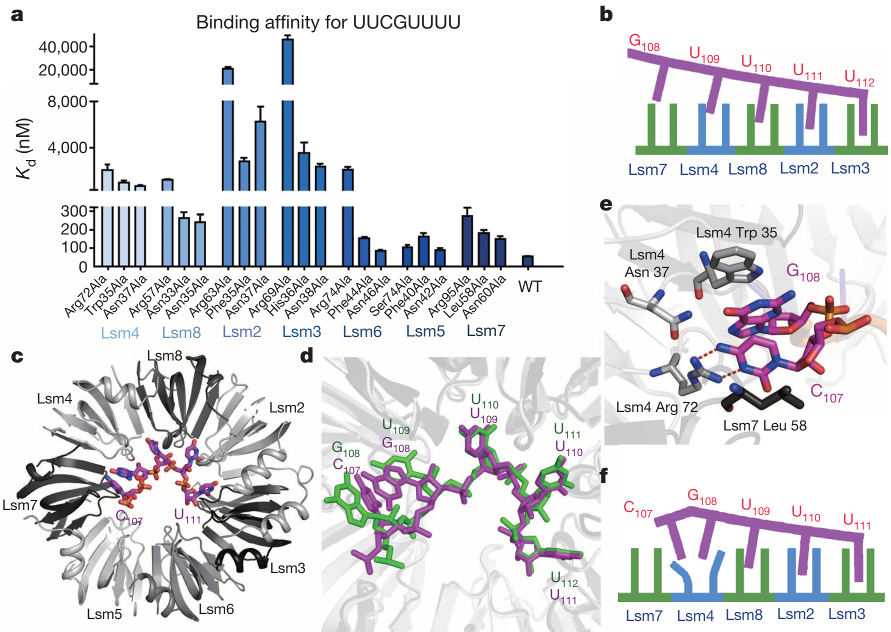


Figure 4 | Lsm3 anchors the 3′ end of RNA elements. **a**, Differential contribution to U6 snRNA recognition by the Lsm components. 21 Lsm heptameric complexes, each containing a missense mutation targeting the base-stacking residues or the conserved Asn, were examined for binding to the octanucleotide 5′-UUCGUUUU$_{112}$-3′. WT, wild type. **b**, A proposed explanation for why the Lsm complex only accommodates four nucleotides. **c**, Structure of the Lsm2–8 heptameric complex bound to the RNA fragment 5′-UUUCGUUU$_{111}$-3′. **d**, Structural comparison of the Lsm complexes bound to 5′-UUUCGUUU$_{111}$-3′ and 5′-UUCGUUUU$_{112}$-3′. **e**, A close-up view on the accommodation of the dinucleotide C$_{107}$-G$_{108}$ by the same general location as that for U$_{109}$ in the wild-type complex. **f**, A cartoon representation of the recognition of the RNA fragment 5′-UUUCGUUU$_{111}$-3′ by the Lsm2–8 complex.

The 3′ end recognition of U6 snRNP is unique among all RNA-binding proteins, with Lsm3 playing an essential role in anchoring the 3′-uridine of U$_{112}$. To disrupt 3′ end recognition by Lsm3, we deleted U$_{112}$, crystallized the Lsm2–8 complex with the RNA 5′-UUUCGUU U$_{111}$-3′, and solved the structure at 2.6 Å resolution (Fig. 4c and Extended Data Table 1). We had anticipated that this design might result in the binding of G$_{108}$-U$_{109}$-U$_{110}$-U$_{111}$ to Lsm7-Lsm4-Lsm8-Lsm2, with U$_{111}$ anchored by Lsm2 instead. To our surprise, the 3′ end uridine of U$_{111}$ is still anchored by Lsm3 in exactly the same manner as for U$_{112}$ (Fig. 4d). In the structure, U$_{110}$ and U$_{109}$ are recognized by Lsm2 and Lsm8, respectively. Intriguingly, G$_{108}$ and C$_{107}$ are now accommodated in the same general location as U$_{109}$ of the previous structure, prying open the two blades of the sandwich — Trp 35 and Arg 72 from Lsm4 (Fig. 4e, f). This analysis reveals a striking ability for Lsm3 to anchor the 3′ end of bound RNA elements.

The different modes of RNA recognition — end recognition versus internal RNA recognition — seem to match perfectly the different functions and assembly kinetics between the Lsm and Sm complexes. Unlike other snRNAs, U6 snRNA remains constitutively in the nucleus; the Lsm2–8 complex is pre-assembled before recognition of U6 snRNA in the nucleus[25]. By contrast, assembly of other snRNPs occurs in the cytoplasm, where the heptameric Sm complex can only be assembled in the presence of relevant snRNA[25]. Thus, the various snRNA may serve as a nucleation centre for assembly of the Sm–snRNA complex in the cytoplasm. Assembly of the Sm ring *in vivo* is facilitated by chaperone proteins such as the SMN complex[26]. The La protein is required for the maturation of U6 snRNA[27]. Similar to the Sm ring, assembly of the Lsm ring could also be facilitated by other yet-to-be identified chaperones. Our structural revelations, together with biochemical analyses, serve as an important framework for mechanistic understanding of the U6 snRNP function in pre-mRNA splicing.

METHODS SUMMARY

The seven proteins Lsm2–8 were individually cloned into the pQLink vector[28] and co-expressed in *Escherichia coli*, purified to homogeneity, and crystallized by the hanging-drop vapour-diffusion method. Much protein engineering effort was directed at obtaining the heptameric complex and improving the quality of crystals. Diffraction data were collected at Shanghai Synchrotron Radiation Facility beamline BL17U and SPring-8 beamline BL41XU and processed with HKL2000[29]. The crystals of RNA-free Lsm2–8 complex belong to the space groups $I2_12_12_1$ and $P2_1$. The structure was determined at 2.8 Å resolution by molecular replacement using the atomic coordinates of the Lsm1–7 heptameric complex (PDB code 4M75). The Lsm2–8 heptameric complex bound to an octanucleotide derived from the 3′ end of U6 snRNA was crystallized in the space group *C*2. The structure was determined by molecular replacement at 2.8 Å resolution and refined with PHENIX[30]. Dissociation constants for interactions between the Lsm2–8 complex and various U6 snRNA fragments were determined by isothermal titration calorimetry (ITC).

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 August; accepted 29 October 2013.
Published online 17 November 2013.

1. Moore, M. J., Query, C. C. & Sharp, P. A. In *The RNA World* (ed. Atkins, R. G. J.) 303–357 (Cold Spring Harbor Laboratory Press, 1993).
2. Hinterberger, M., Pettersson, I. & Steitz, J. A. Isolation of small nuclear ribonucleoproteins containing U1, U2, U4, U5, and U6 RNAs. *J. Biol. Chem.* **258,** 2604–2613 (1983).
3. Fabrizio, P., Esser, S., Kastner, B. & Luhrmann, R. Isolation of *S. cerevisiae* snRNPs: comparison of U1 and U4/U6.U5 to their human counterparts. *Science* **264,** 261–265 (1994).
4. Will, C. L. & Luhrmann, R. In *The RNA World* 3rd edn (eds Gesteland, R.F., Cech, T. R. and Atkins, J. F.) 181–204 (Cold Spring Harbor Laboratory Press, 2006).
5. Achsel, T. *et al.* A doughnut-shaped heteromer of human Sm-like proteins binds to the 3′-end of U6 snRNA, thereby facilitating U4/U6 duplex formation *in vitro*. *EMBO J.* **18,** 5789–5802 (1999).
6. Mayes, A. E., Verdone, L., Legrain, P. & Beggs, J. D. Characterization of Sm-like proteins in yeast and their association with U6 snRNA. *EMBO J.* **18,** 4321–4331 (1999).
7. Cooper, M., Johnston, L. H. & Beggs, J. D. Identification and characterization of Uss1p (Sdb23p): a novel U6 snRNA-associated protein with significant similarity to core proteins of small nuclear ribonucleoproteins. *EMBO J.* **14,** 2066–2075 (1995).
8. Black, D. L. & Steitz, J. A. Pre-mRNA splicing *in vitro* requires intact U4/U6 small nuclear ribonucleoprotein. *Cell* **46,** 697–704 (1986).

9. Berget, S. M. & Robberson, B. L. U1, U2, and U4/U6 small nuclear ribonucleoproteins are required for *in vitro* splicing but not polyadenylation. *Cell* **46,** 691–696 (1986).
10. Datta, B. & Weiner, A. M. Genetic evidence for base pairing between U2 and U6 snRNA in mammalian mRNA splicing. *Nature* **352,** 821–824 (1991).
11. Wu, J. A. & Manley, J. L. Base pairing between U2 and U6 snRNAs is necessary for splicing of a mammalian pre-mRNA. *Nature* **352,** 818–821 (1991).
12. Yean, S.-L., Wuenschell, G., Termini, J. & Lin, R.-J. Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome. *Nature* **408,** 881–884 (2000).
13. Brow, D. A. & Guthrie, C. Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. *Nature* **334,** 213–218 (1988).
14. Jandrositz, A. & Guthrie, C. Evidence for a Prp24 binding site in U6 snRNA and in a putative intermediate in the annealing of U6 and U4 snRNAs. *EMBO J.* **14,** 820–832 (1995).
15. Martin-Tumasz, S., Reiter, N. J., Brow, D. A. & Butcher, S. E. Structure and functional implications of a complex containing a segment of U6 RNA bound by a domain of Prp24. *RNA* **16,** 792–804 (2010).
16. Martin-Tumasz, S., Richie, A. C., Clos, L. J. II, Brow, D. A. & Butcher, S. E. A novel occluded RNA recognition motif in Prp24 unwinds the U6 RNA internal stem loop. *Nucleic Acids Res.* **39,** 7837–7847 (2011).
17. Leung, A. K., Nagai, K. & Li, J. Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. *Nature* **473,** 536–539 (2011).
18. He, W. & Parker, R. Functions of Lsm proteins in mRNA degradation and splicing. *Curr. Opin. Cell Biol.* **12,** 346–350 (2000).
19. Zaric, B. *et al.* Reconstitution of two recombinant LSm protein complexes reveals aspects of their architecture, assembly, and function. *J. Biol. Chem.* **280,** 16066–16075 (2005).
20. Kambach, C. *et al.* Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96,** 375–387 (1999).
21. Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K., Li, J. & Nagai, K. Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature* **458,** 475–480 (2009).
22. Weber, G., Trowitzsch, S., Kastner, B., Luhrmann, R. & Wahl, M. C. Functional organization of the Sm core in the crystal structure of human U1 snRNP. *EMBO J.* **29,** 4172–4184 (2010).
23. Lund, E. & Dahlberg, J. E. Cyclic 2′,3′-phosphates and nontemplated nucleotides at the 3′ end of spliceosomal U6 small nuclear RNA's. *Science* **255,** 327–330 (1992).
24. Licht, K., Medenbach, J., Luhrmann, R., Kambach, C. & Bindereif, A. 3′-cyclic phosphorylation of U6 snRNA leads to recruitment of recycling factor p110 through LSm proteins. *RNA* **14,** 1532–1538 (2008).
25. Raker, V. A., Hartmuth, K., Kastner, B. & Luhrmann, R. Spliceosomal U snRNP core assembly: Sm proteins assemble onto an Sm site RNA nonanucleotide in a specific and thermodynamically stable manner. *Mol. Cell. Biol.* **19,** 6554–6565 (1999).
26. Chari, A. *et al.* An assembly chaperone collaborates with the SMN complex to generate spliceosomal SnRNPs. *Cell* **135,** 497–509 (2008).
27. Pannone, B. K., Xue, D. & Wolin, S. L. A role for the yeast La protein in U6 snRNP assembly: evidence that the La protein is a molecular chaperone for RNA polymerase III transcripts. *EMBO J.* **17,** 7442–7453 (1998).
28. Scheich, C., Kummel, D., Soumailakakis, D., Heinemann, U. & Bussow, K. Vectors for co-expression of an unrestricted number of proteins. *Nucleic Acids Res.* **35,** e43 (2007).
29. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
30. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).

Acknowledgements We thank S. Huang and J. He at SSRF beamline BL17U and N. Shimizu, T. Kumasaka, and S. Baba at the Spring-8 beamline BL41XU for on-site assistance. This work was supported by funds from National Natural Science Foundation of China projects 31130002 and 31021002.

Author Contributions L.Z., J.H., and Y.S. designed all experiments. L.Z., J.H., Y.Z., R.W., G.L., P.Y., and C.Y. performed the experiments. All authors contributed to data analysis. L.Z., J.H., C.Y., and Y.S. contributed to manuscript preparation. Y.S. wrote the manuscript.

Author Information Atomic coordinates and structure factors have been deposited in the Protein Data Bank. The PDB codes of RNA-free Lsm2–8 complex are 4M77 and 4M78 for space groups $I2_12_12_1$ and $P2_1$, respectively. The PDB codes of Lsm2–8 bound to the RNA elements 5′-UUCGUUUU-3′ and 5′-UUUCGUUU-3′ are 4M7A and 4M7D, respectively. The PDB code of Lsm1–7 is 4M75. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.S. (shi-lab@tsinghua.edu.cn).

CAREERS

METRICS Blog references to papers linked to more journal citations **p.123**

STUDENTS Doctoral scholarships offered in United Kingdom **p.123**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

ILLUSTRATION BY DAVID SIMONDS

TELECOMMUTING

No place like home

Researchers can avoid stressful commutes and boost efficiency by working from home.

BY KAREN KAPLAN

Peter Griffith is not a fan of traffic in the Washington DC area. On a bad day, his 103-kilometre round-trip commute means that he can spend more than four hours on the road. By the time he gets to his desk at NASA's Goddard Space Flight Center, Griffith — like millions of metropolitan commuters worldwide — feels drained.

Griffith, who is chief support scientist for NASA's carbon cycle and ecosystems office in Greenbelt, Maryland, coordinates the North American Carbon Program at Goddard. His work involves computer analysis of remote-sensing and geospatial data, for which he needs long, uninterrupted blocks of time — almost impossible when the phone is ringing, e-mails are pouring into his in-box and people are knocking on his door. So to save commuting time and ensure solitude, he works from home twice a week. "I don't have a wet lab. I don't have an engineer," says Griffith. "I'm not one of the people at Goddard building a satellite. It's just easier for me to do a lot of my work from home."

Griffith is one of a growing number of scientists around the world who are enjoying the benefits of working from home. According to the US Census Bureau, about 13.4 million people in the United States worked from home for at least one day a week in 2010, up 41% from a decade earlier.

The practice is not new for researchers, who have long worked from home writing grant applications and research papers, grading exams or preparing lectures. But advances in technology are facilitating and accelerating this trend, allowing researchers to do more from home than ever before, especially if they work in bioinformatics or computational science (see *Nature* **504,** 319–321; 2013). An Internet connection provides access to everything from e-mail to remote supercomputers; Skype and other computer programs enable inexpensive long-distance voice and video chats; and applications such as Google Drive allow multiple users to access documents remotely and simultaneously.

Early-career researchers who want to work from home will need to determine how to balance their obligations. A researcher may be required to develop a written proposal for their principal investigator, supervisor or ▶

department head that explains why they want to work from home, the work they will do and their projected home-based schedule.

At the very least, the researcher should be prepared to draw up talking points for a discussion of the proposal. "You need to have a clear story," emphasizes Ferdinand Grozema, a chemist at the Delft University of Technology in the Netherlands.

Early-career researchers should also clarify how they plan to address problems that may arise. For example, they may need to explain how a glitch with an experiment they are managing might be handled in their absence, or how they can attend a lab or department meeting virtually. And, of course, to work most successfully from home generally means that one should be conducting the type of research — like Griffith's — that does not require eight to ten hours in the lab each day (see 'Comforts of home').

BREATHING SPACE

Researchers who work regularly from home cite quiet time and the absence of disruption as the primary benefits.

Paul Bédard, a geochemist at the University of Quebec in Chicoutimi, Canada, spends one day a week at home processing data sets. While there, he also prepares course material for classes that he teaches in mineralogy and geostatistics, and works on grant applications and papers. "If I'm at the office, I am constantly getting a knock on my door from students or colleagues," he says. "You need quiet time for more than a few minutes to do this work, and at home I have a few hours. You need breathing and thinking space — you need to let your brain wander around. That's where you find the solution, the answer."

Alison Diaper, who juggles jobs as a contract researcher in mental health and addiction at the University of Bristol, UK, and as a clinical-trials manager at Frenchay Hospital in Bristol, says, "I can escape random questions, other colleagues and the telephone ringing." She works from home once a week or so, using the time to set up studies, analyse data and write up results for both jobs.

There are also other practical considerations. Bédard is happy to avoid the commute through Quebec's wintry climes. "I don't have to drive to the university during a big snowstorm," he says.

Marcel Swart, a theoretical chemist at the University of Girona in Spain, likes to avoid the tourist traffic in the summer that swarms in from the coast, east of his home in La Bisbal d'Empordà. "They don't know where they're going," he says. "You can't go more than 30 kilometres an hour."

SETTING UP

Aspiring scientist telecommuters need to notify managers, lab mates, colleagues and students of their home schedule well in advance. Researchers who are used to working remotely say that their regular notification routine includes sending out e-mails and texts, leaving voicemail messages and posting notices on their lab calendars and office doors at the beginning of the week — in some cases, up to ten days in advance — with their home-based schedule and contact information.

If necessary, an early-career researcher should make it clear to colleagues, including managers and graduate students, that calls, texts and e-mails will receive responses only at certain times of the day. But veteran home workers say that it is crucial to have consistent access to colleagues, especially for recently appointed faculty members.

Catherine Cardelús, a biologist at Colgate University in Hamilton, New York, has 6–12 undergraduate students in her lab all year. When Cardelús is out of the lab, she ensures that everyone knows her schedule and that she is available by phone or online. "I want research constantly done, so I make sure my students have what they want and need," she

TIPS FOR WORKING REMOTELY

Comforts of home

To be effective at working from home, keep these guidelines in mind.

- Determine which tasks are best done at the workplace. Working remotely using screen-sharing software often changes the dynamics of a collaboration, and team members should clarify who has the final say on changes, or try to produce final drafts in person.
- Set up your teleworking schedule to overlap as much as possible with those of your lab head, supervisor, colleagues and anyone else with whom you regularly interact.
- Determine what portion of your work is best handled by e-mail or online versus by phone. For detailed calculations that require input from colleagues, for example, e-mail is best. A written record helps to minimize errors and misunderstandings.
- Keep on top of tasks that need to be done in person at the lab or office, such as taking measurements or signing paperwork.
- Arrange regular coffees and lunches with colleagues and others while at the lab or office to catch up on informal workplace news exchanges. **K.K.**

LES TODD/DUKE PHOTOGRAPHY

“I’m not one of the people at Goddard building a satellite. It’s just easier for me to do a lot of my work from home.”
Peter Griffith

says. “If you want an active lab, you have to be accessible. My students can text me with questions such as, ‘When are you going to be back in the lab?’ or ‘How do we order some HCl?’. I make sure that what they need is always there, and that’s what has allowed me to work at home when I do.” If she needs to stay off e-mail or her mobile for an hour or two, she does so, but provides ample warning that she will be unavailable.

Getting used to providing an open line of communication and a transparent schedule may be an adjustment for researchers who have been accustomed to more autonomy, she warns. “The biggest shocker for most early-career faculty members is how hard it is to be able to stay at home because people rely on you to be in your lab and your office.”

Depending on the institution, there may be thorny or murky policy issues on telecommuting to contend with. When Grozema’s first child was born and he wanted to work from home, he elected to take a day’s paternity leave per week for about one-third less pay for that day.

But when his second child arrived about a year ago, and Grozema considered working from home again, he discovered that many of his colleagues regularly worked from home without having to take leave and get paid less — the policy was not well defined. He approached his department head, and the two worked out an agreement under which Grozema uses a half-day’s leave per week when he works from home.

Once remote workers have settled on a schedule, they need to stick to it, say researchers. If time at home provides the luxury of several hours without interruption, an early-career researcher needs to use that time to actually do work — many warn that it is all too easy to give in to the siren song of smartphones and social media. “You have to motivate,” says Diaper. “You have to be strict and say to yourself that you have to get the job done. You can’t be swayed by your partner’s request or your own temptation.”

DEALING WITH DOWNSIDES

There are other pitfalls for those who work from home, including the possibility of a lower profile because of reduced visibility. Cardelús says that it is wise to interact regularly and often in person with colleagues, associates and superiors. Working from home “can be very isolating”, she says. “You need to be networking — you need to be seen.”

Some ways of counteracting the potential ‘out of sight, out of mind’ problem include securing a mentor who is particularly sympathetic to junior researchers’ telecommuting and career-support needs. An understanding mentor might help to keep a home worker’s profile high by routinely talking up their work, thus mitigating the impact of decreased visibility.

People who work from home do risk missing impromptu chats, which can do more than just provide entertainment or build rapport — they offer access to unofficial intelligence that is a key part of understanding the changing dynamics of every workplace. “When I’m home, I miss out on going to have coffee with people, and that’s when all kinds of information about employment applications, the ministries and the university comes up,” says Swart. “If I’m not there, I don’t go out — and this kind of information is never shared on e-mail.”

“I can escape random questions, other colleagues and the telephone ringing.”
Alison Diaper

It can be a challenge, lament early-career researchers, to deal with the lack of connection with lab mates and department colleagues. “You don’t have people around you to talk about what you’re doing,” says Richard Lonsdale, a postdoc in computational chemistry at the University of Bristol, UK. Lonsdale, who lives two hours’ drive away in Birmingham, has been working from home for four days per week, and makes sure that he regularly e-mails colleagues and sets up Skype chats to confer about ideas when he is at home. He also arranges in-person discussions and meetings for days on which he comes in to the university. “You have to make the most of the day when you’re in the lab,” he says.

Scientists who routinely work from home agree that it takes effort to counterbalance the downsides. But that is not a deal-breaker, they say. “It’s not unpleasant to be at a bit of a distance,” says Grozema, who adds that a day of telecommuting per week has helped with his work–life balance. “You don’t have to be less productive.” ■

Karen Kaplan *is the associate Careers editor at* Nature.

METRICS

Blog citations count

Papers that are formally cited by research-oriented blogs receive more journal citations, finds a study published on 15 January (H. Shema *et al. J. Assoc. Inf. Sci. Technol.* http://doi.org/q88; 2014). For 7 of the 12 scientific journals examined in 2009, and 13 of 19 journals analysed in 2010, papers cited in blog posts aggregated by ResearchBlogging.org received more subsequent citations than did papers from the same journal in the same year that had not been cited by blogs. Hiring and tenure-review committees could use blog citations to assess the impact of recently published papers, suggests co-author Hadas Shema, an information scientist at Bar-Ilan University in Ramat-Gan, Israel.

TRAINING

Doctorates diversify

Leading European Union (EU) research universities are adding career development to their doctoral programmes, including schemes to help postgraduates into non-academic careers, finds a 27 January report by the League of European Research Universities (LERU) in Leuven, Belgium. Institutions are increasingly offering options including employer-led career-skills workshops, employment forums and fairs, student consultancies and internships with industry, it found. A LERU report four years ago called for such expansion in the face of declining academic research positions and a tight economic climate. Doctoral students sometimes do not appreciate the rare number of academic posts, and institutions need to offer guidance for alternatives, says Katrien Maes, LERU’s chief policy officer.

SCHOLARSHIPS

Trust funds PhDs

The Leverhulme Trust, a non-profit research funder in London, will invest £10 million (US$16.6 million) to create 150 doctoral scholarships across all UK science and humanities disciplines. Each award will be for £70,000 over 36 months. Universities can opt to offer extra funding to awardees, says trust spokesman Daniel Mapp. The scheme is meant to help those with undergraduate debt to pursue PhD degrees, but winners do not have to aim for any one professional path. “It will be for individuals to decide how they take their careers forward,” Mapp says. Anyone at a UK university is eligible, but UK and European Union students get priority.

VESSELS FOR DESTRUCTION

The price of perseverance.

BY A. G. CARPENTER

Muhmughmuhmuh. The hushed babble of the spectators crests as the guards bring the prisoner into the hall.

Patron Jamis looks up at the gallery, stern, and the mutters fade.

The girl looks like all the rest of her kind, dirty and scarred, but she has the nine bands of Authority tattooed on her forehead and she stands straight, even under the weight of her chains.

Jamis clasps his hands behind his back, considering. They've had other Destructives before, but never one with the full nine. It means she is a leader and a prophet. Maybe even *the* leader and prophet.

He clears his throat.

"We have ways of making you talk."

She grins. "We have ways of making you talk." The words are sing-song.

One of the guards steps forward, truncheon raised, but the girl doesn't flinch and Jamis shakes his head, motioning the guard away.

"It will be more pleasant for everyone if you cooperate."

She shrugs, awkward with her arms bound to the pole across her shoulders. "I am cooperating. I let you take me in."

"You were gravely outnumbered."

"Because I chose to be." A quick shake of her head and the irritated quirk in her mouth relaxes. "Tell me what you want to know."

Jamis frowns. This feels wrong. Easy. Not at all like he has anticipated.

"Well?" she asks.

"They say you can see the future."

"Yes."

"Yes that's what they say, or yes you can?"

She raises an eyebrow. "Yes, I can."

The spectators take a collective breath; the noise breaks on the high curve of the ceiling like water and falls back in bits and pieces. *Ahhh-ahhs-sh-shs.*

"Then you must know why I have brought you here."

"I do." A pause. "Oh. You want me to tell you?"

"Yes, I want you to tell me." His voice is rough with annoyance.

"You want me to tell you if you will succeed in wiping out the Destructives."

Ahhh-ahhs-sh-shsh.

Jamis waits for the echoes to fade. "Yes."

Another shrug. "Of course."

His heart hammers in his chest.

"Of course?"

"You don't believe me."

JACEY

"How can you be certain of this?"

She looks at him and there is the hint of deep water in her eyes. Fathomless. Calm. Dangerous. "I have seen the future of mankind, Patron Jamis. It is death."

"We will persevere to the bitter end." It is an automatic response.

"Of course you will." She coughs, licks the spittle from her lips, wincing when her tongue finds a bloodied split.

The spectators rumble, uneasy.

Something cold and prickly writhes in Jamis's gut. Guilt or fear or anger, he can't tell.

The girl leans closer, something like pity making lines around her mouth. "It's okay. You can't help it. None of us can."

"Everyone wants to survive."

She shakes her head. "Everyone wants to die."

Muhmughmuhmuh.

He raises a hand and the spectators hush. "You speak lies."

"It is a part of our DNA. The Romans lined copper cups with lead because it made the wine taste sweeter. The European settlers in America farmed tobacco. You created vaccines for every ailment known to man, till your bodies could no longer defend against a live disease."

"We didn't know any better at the time."

"No." She sighs. "We cannot help it."

His gut twinges, but he is not one of her disciples. "We will survive."

She says nothing.

"We are not mere animals. We were made for more than that."

"Our souls, you mean." She smiles, sunlight on desert sand. "Maybe. But this…" She waggles her fingers. "This is just a vessel built for destruction."

The room is like a bell, clamouring with the uncertainty of the crowd in the gallery.

Jamis silences it with a single gunshot. It is a shame to waste a bullet on the girl when a knife would have done just as well, but it would not have caught their attention in the same way.

He tucks the pistol back into its holster. "Take that out."

The guards drag the body away and Jamis turns to face the gallery.

Hundreds of eyes burn at him, candle flames flickering against the dark nothingness.

"If it is destruction they want, they shall have it."

They sigh in agreement.

Yessssssssssss.

"We will fight against this seed of despair until it is wiped from the face of the Earth."

Yessssssssssss.

"Go and burn them out."

Thunder shakes the room, feet pounding the metal floor as they trample out of the doors. Men, women and children all bent to one purpose.

Jamis returns to his room. His face itches and he scratches it, absent. His fingers come away speckled with blood. Something in his gut stirs, uncomfortable. Even in death the girl is filthy.

He washes and waits for the army to return.

Three days later they find him in his room, bright-eyed with fever. His cheeks are spotted with black. His hands, too.

The guards draw back, a single word on their breath.

Pox.

Jamis nods and smiles, mad with certainty. "I have seen the future." He coughs, lips wet with spittle. "It is death." ■

A. G. Carpenter *writes speculative fiction of (and for) all sorts. Her work has appeared in* Daily Science Fiction, Abyss & Apex *and* Stupefying Stories. *She lives in the southeastern United States.*